GHEORGHE MIHOC
VIRGIL CRAIU

TRATAT DE STATISTICĂ MATEMATICĂ

VOLUMUL IV

CORELAȚIE ȘI REGRESIE LINIARĂ

EDITURA ACADEMIEI REPUBLICII SOCIALISTE ROMÂNIA

Scanned with OKEN Scanner

GHEORGHE MIHOC ◆ VIRGIL CRAIU

TRATAT de STATISTICĂ MATEMATICĂ

Volumul IV

CORELAȚIE ȘI REGRESIE LINIARĂ

Scanned with OKEN Scanner

PREFAȚĂ

Din cele mai vechi timpuri, oamenii au fost interesați în problema alegerii unei singure valori care să caracterizeze informația dată de un număr de observații sau măsurători independente, fiecare supusă la erori. Grecii și egiptenii vechi și arabii în evul mediu foloseau modulul. Mediana și media aritmetică au fost și sînt folosite și în prezent.

Problema determinării coeficienților ecuației liniei drepte care ajustează cel mai bine (într-un sens specificat) trei sau mai multe puncte necoliniare în planul (xoy) ale căror coordonate sînt perechile de valori asociate cu două variabile X și Y datează, după cunoștința noastră, de la Galileo Galilei (1632). Această problemă poate fi generalizată în două moduri: i) în loc de a ajusta cea mai bună ecuație în două variabile (cea mai bună linie în plan), putem găsi cea mai bună ecuație în trei variabile (cel mai bun plan în spațiu cu trei dimensiuni) sau cea mai bună ecuație în mai multe variabile (cel mai bun hiperplan într-un hiperspațiu), ii) renunțăm la cerința ca ecuația să fie liniară și găsim cea mai bună curbă în plan, cea mai bună suprafață în spațiu cu trei dimensiuni sau cea mai bună hipersuprafață în hiperspațiu. Ne vom referi la aceste probleme ca la regresia liniară și regresia neliniară.

Problemele determinării celei mai bune măsuri a tendinței centrale și a celei mai bune ecuații a regresiei liniare sau neliniare sînt legate una de alta și de problema alegerii celei mai bune măsurători a dispersiei. Soluțiile tuturor acestor trei probleme depind de repartiția erorilor sau reziduurilor (abaterile valorilor observate de la acelea predictate de ecuația de regresie). Dacă presupunem că valorile variabilei independente sînt cunoscute exact și că erorile în observațiile asupra variabilelor dependente sînt repartizate normal, atunci media este cea mai bună măsură a tendinței centrale, abaterea standard este cea mai bună măsură a dispersiei și metoda celor mai mici pătrate este cea mai bună metodă pentru ajustarea ecuației de regresie. Alte presupuneri asupra erorilor conduc la alegeri diferite. Mulți practicieni au tendința de a face presupunerea de normalitate și nu examinează consecințele cînd această presupunere nu este justificată. O altă problemă apare

Scanned with OKEN Scanner

cînd datele sînt contaminate de observaţii ce diferă mult de celelalte date (date nesemnificative), date care provin din repartiţii cu medii diferite şi/sau abateri standard mari. Diferitele metode pentru tratarea datelor nesemnificative sînt prezentate de Craiu, Bad (1979).

În acest volum am încercat să prezentăm cele mai reprezentative aspecte din teoria corelaţiei şi regresia liniară, urmînd că în celelalte volume să ne ocupăm de regresia neliniară, analiza dispersională, teoria deciziilor şi analiza Bayesiană, şi de seriile dinamice.

AUTORII

Scanned with OKEN Scanner

CUPRINS

Capitolul XXVIII

COEFICIENȚI DE CORELAȚIE

§ 1. Repartiția normală bidimensională 19
§ 2. Coeficientul de corelație de selecție R. Repartiția statisticii R în populații normale 28
§ 3. Repartiția statisticii R în populații care nu sînt normale 33
§ 4. Aproximații și tabele 34
§ 5. Covarianța de selecție 41
§ 6. Corelație serială circulară 43
§ 7. Corelație serială necirculară 46
§ 8. Repartiția Leipnik 51
§ 9. Coeficientul de corelație multiplă 54
§ 10. Repartiția normală tridimensională. Corelație parțială 61
§ 11. Variabile dichotomizate 69
11.1. Corelație tetrachorică 69
11.2. Corelația biserială 72
§ 12. Repartiții normale bidimensionale și tridimensionale trunchiate 74
§ 13. Corelații canonice 77
13.1. Populații normale 77
13.2. Populații care nu sînt normale 79
§ 14. Coeficientul de corelație pentru date direcționale. 82
§ 15. Corelația rangurilor 89
15.1. Coeficientul lui Spearman 89
15.2. Coeficientul T al lui Kendall 93
§ 16. Alte măsuri ale dependenței 94
16.1. Extinderea noțiunii de coeficient de corelație la variabile oarecare 94
16.2. Un coeficient de corelație global 94
16.3. Corelația informațională și coeficientul de corelație informațională 96
16.3.1. Repartiții discrete 96
16.3.2. Repartiții cu densitatea continuă. 98
16.4. Corelație și coeficienți de corelație entropică 100

Capitolul XXIX

INFERENȚE PRIVIND COEFICIENȚII DE CORELAȚIE ȘI MATRICILE DE COVARIANȚE

§ 1. Estimarea parametrilor repartiției normale bidimensionale. . . . 105
1.1. Selecție totală (nedesfăcută) 105
1.2. Estimarea coeficientului de corelație dintr-o selecție aleatoare desfăcută 110

Scanned with OKEN Scanner

1.2.1. Estimarea parametrilor necunoscuți 110
1.2.2. Funcția de verosimilitate integrată 112
1.2.3. Informația lui Fisher 114
§ 2. Inferențe privind coeficienții de corelație cu ajutorul aproximației lui Fisher 115
2.1. Test pentru un ρ dat 115
2.2. Test pentru egalitatea a doi coeficienți de corelație 116
2.3. Test pentru omogenitatea unei mulțimi de coeficienți de corelație 117
§ 3. Teste de independență în repartiția normală bidimensională . . . 117
§ 4. Verificarea ipotezei de independență pentru date direcționale 119
§ 5. Teste privind coeficientul de corelație multiplu și coeficientul de corelație parțial multiplu 120
5.1. Introducere 120
5.2. Invarianți maximali și teste ale raportului de verosimilități 121
5.3. Repartiția statisticilor Z_2 și Z_3 123
5.4. Teste pentru problemele i)—iv) 125
§ 6. Teste în analiza corelației canonice pentru populații care nu sînt normale 126
§ 7. Teste pentru ipoteze asupra matricilor de covarianțe pentru populații care nu sînt normale 130
7.1. Test pentru verificarea ipotezei conform căreia matricea de covarianțe este egală cu o matrice dată 130
7.2. Teste pentru sfericitate 132
7.3. Test pentru egalitatea a două matrici de covarianțe 132
7.4. Test pentru verificarea ipotezei conform căreia Σ este o matrice diagonală bloc 133

Capitolul XXX

VALORI MEDII CONDIȚIONATE. REGRESIE

§ 1. Curbe de regresie 135
§ 2. Regresie liniară. Coeficienți de regresie 136
§ 3. Coeficientul de corelație 138
§ 4. Raport de corelație 140
§ 5. Legătura dintre curba de regresie și cumulanți 141
§ 6. Regresii liniare parțiale 149
§ 7. Coeficienții de regresie de selecție 153
§ 8. Coeficienții de selecție 156
§ 9. Teste pentru rapoartele de corelație și liniaritatea regresiei 158
§ 10. Aplicații ale proprietății de nedeplasare la regresie 160
§ 11. Modele ale regresiei pentru date direcționale 162

Capitolul XXXI

REGRESIE LINIARĂ. MODELUL DE RANG MAXIM

§ 1. Definiția modelului liniar general. Presupuneri 164
§ 2. Estimarea parametrilor 165
2.1. Metoda C.M.M.P. 165
2.2. Proprietățile estimatorului de C.M.M.P. pentru β 169
2.3. Sume de pătrate. Estimarea lui σ^2 171

Scanned with OKEN Scanner

2.4. Repartiția estimatorului lui β. Repartiții ale sumelor de pătrate 174
2.5. Eroarea pură 177
2.6. Verificarea ipotezelor 178
2.7. Intervale de încredere 179
§ 3. Ipoteza liniară generală 181
3.1. Verificarea ipotezei liniare generale 181
3.2. Estimație în ipoteza nulă 183
3.3. Aplicații 184
§ 4. Modele reduse 187
§ 5. Examinarea reziduriilor 192

Capitolul XXXII

MODELE CARE NU SÎNT DE RANG MAXIM

§ 1. Modelul. Presupuneri 195
§ 2. Estimarea parametrilor 195
2.1. Ecuațiile normale 195
2.2. Proprietățile soluției ecuațiilor normale 196
2.3. Sume de pătrate. Estimarea lui σ^2 197
2.4. Repartiția soluției ecuațiilor normale. Repartiții ale sumelor de pătrate 199
2.5. Verificarea ipotezelor 202
§ 3. Funcții estimabile 203
3.1. Definiția funcției estimabile 203
3.2. Proprietățile funcțiilor estimabile 204
3.3. Intervale de încredere 206
3.4. Funcții ce sînt estimabile 206
§ 4. Ipoteze liniare generale 209
4.1. Ipoteze verificabile 209
4.2. Ipoteze care nu sînt verificabile 214
§ 5. Modele restrînse 218
5.1. Introducere 218
5.2. Restricții conținînd funcții estimabile 219
5.3. Restricții conținînd funcții care nu sînt estimabile 220
Probleme 221

Capitolul XXXIII

ANALIZA REGRESIEI PRIN RĂDĂCINI CARACTERISTICE

§ 1. Introducere 223
§ 2. Obținerea ecuațiilor de predicție 224
§ 3. Interpretarea geometrică a rădăcinilor și vectorilor proprii ai matricei A'A 226
§ 4. Estimatori de C.M.M.P. și estimatori de C.M.M.P. modificați . . . 227
§ 5. Eliminarea variabilelor independente 229
§ 6. Proprietăți asimptotice 231

Scanned with OKEN Scanner

Capitolul XXXIV

REGRESIA RIDGE. ESTIMATORI PRIN INVERSĂ GENERALIZATĂ

§ 1. Regresia ridge 235
1.1. Introducere 235
1.2. Regresia ridge 236
1.3. Urma ridge 238
1.3.1. Definiția urmei ridge. Caracterizarea urmei ridge 238
1.3.2. Caracterizarea prin verosimilități a urmei ridge 239
1.4. Proprietăți ale mediei erorii pătratice a regresiei ridge 240
1.5. Forma generală a regresiei ridge 240
§ 2. Estimatori ai regresiei prin inversă generalizată 243
2.1. Inverse generalizate 243
2.2. Proprietăți ale soluției inversă generalizată 245
§ 3. O clasă de estimatori contractați 248
3.1. Punerea problemei 248
3.2. Estimatori ridge și generalizările lor 250
3.3. C.M.M.P. cu restricții 252
§ 4. Estimatori ridge buni și optimi 254
4.1. Introducere 254
4.2. Factori ridge buni 256
4.3. Factori ridge optimi 258
§ 5. Validarea mixtă pentru alegerea unui parametru ridge bun . . . 260

Capitolul XXXV

REGIUNI DE ÎNCREDERE DE DIMENSIUNE FIXATĂ PENTRU PARAMETRII REGRESIEI

§ 1. Problema I 266
1.1. Soluția 266
1.2. Teoria asimptotică pentru valori mari ale lui n 271
1.3. Proprietăți asimptotice ale clasei $\mathscr{C}$ 272
§ 2. Problema 2 275
2.1. Soluția 275

Capitolul XXXVI

REGRESIE ROBUSTĂ. ESTIMAȚIE

§ 1. Introducere 279
§ 2. Estimații M 285
2.1. Metoda C.M.M.P. cînd erorile nu sînt normale 285
2.2. Estimații M ale regresiei 287
2.3. Deplasarea 293
2.4. Termenii covarianță 294
2.5. Estimația matricii de covariațe a lui $\hat{\theta}$ 296
2.6. Calcularea unor astfel de estimații 299
2.7. Comportarea asimptotică a estimatorilor M pentru modelul liniar 302
§ 3. Estimații R ale coeficienților de regresie 306
3.1. Construcția unei estimații a parametrului vector de regresie bazată pe alegerea convenabilă a statisticilor de rang 306

Scanned with OKEN Scanner

3.1.1. Notații. Presupuneri 306
3.1.2. Estimarea parametrilor $\Delta_1, \ldots, \Delta_R$ 309
3.1.3. Estimarea parametrului α 310
3.2. Estimarea coeficienților de regresie prin minimizarea dispersiei reziduriilor 311
3.2.1. Definirea unor măsuri ale dispersiei și proprietăți ale estimatorilor regresiei 311
3.2.2. Echivalența asimptotică a estimațiilor propuse de Jurečková și Jaeckel 313
3.3. Estimarea coeficienților de regresie prin minimizarea funcției de dispersie definită de diferența ponderată a lui Gini 317
3.3.1. Definiția și proprietățile funcției de dispersie 317
3.3.2. O clasă generală de variabile independente 320
3.3.3. Liniaritatea asimptotică 323
3.3.4. Repartiția lui $\hat{\beta}$ 327
§ 4. Relații asimptotice ale estimatorilor M și R în modelul liniar al regresiei 332
4.1. Presupuneri și notații 332
4.2. Repartiția asimptotică a statisticii $\hat{\Delta}_M - \hat{\Delta}_R$ 334
4.3. Comportarea asimptotică a lui $M_j(\mathbf{X}_N, \Delta)$ 335
§ 5. Asupra estimării densității de repartiție a erorilor în regresie . . . 335
5.1. Introducere 335
5.2. O estimație a densității de repartiție a erorii 337
5.3. Cazul uniform (rectangular) al funcției de pondere. 339
5.4. O extensie la regresia multiplă 341
5.5. Convergența aproape sigură 342

Capitolul XXXVII

REGRESIE ROBUSTĂ. VERIFICĂRI DE IPOTEZE

§ 1. Teste de rang pentru parametrii mai multor drepte de regresie . . . 344
1.1. Teste de rang bazate pe rangurile individuale a k selecții diferite pentru paralelismul dreptelor de regresie 344
1.1.1. Notații. Presupuneri 344
1.1.2. Teste parametrice asimptotic optime 346
1.1.3. Teste de rang asimptotic cel mai puternice 348
1.1.4. Test pentru un model mai general 350
1.1.5. Monotonia statisticilor de rang de regresie 353
1.2. Teste de rang pentru identitatea intercepțiilor mai multor drepte de regresie 353
1.2.1. Notații. Presupuneri 353
1.2.2. Teste parametrice optime 356
1.2.3. Teste de rang de aliniere 357
1.2.4. Eficiența asimptotică relativă 358
1.3. Teste de rang bazate pe rangurile simultane ale tuturor observațiilor pentru identitatea mai multor drepte de regresie 358
1.3.1. Notații și preliminarii 358
1.3.2. Repartiția asimptotică a statisticii $\hat{L}_n$ 360
1.3.3. Eficiența asimptotică 361
§ 2. Teste de rang pentru subipoteze în regresia liniară generală . . . 362
2.1. Statisticile test de rang și semn 362

Scanned with OKEN Scanner

2.2. Statistici test de rang 365
2.3. Eficiența asimptotică relativă 367
§ 3. Estimarea parametrului de locație după un test preliminar asupra regresiei 369
3.1. Cazul unidimensional 369
3.1.1. Introducere 369
3.1.2. Estimatorii locației 369
3.1.3. Repartiția asimptotică a estimatorului θ_n^* 372
3.2. Cazul multidimensional 375
3.2.1. Introducere 375
3.2.2. Notații preliminare și estimatori 375
3.2.3. Repartiția asimptotică a estimatorilor 377
3.2.4. Deplasarea asimptotică și matricea de covariațe a estimatorilor 379
3.2.5. Compararea estimatorilor 380
§ 4. Convergența tare a statisticilor de rang de regresie 381
4.1. Introducere 381
4.2. Statisticele de regresie ale semnului rangului 385

Capitolul XXXVIII

REGRESIE ROBUSTĂ. PROCEDEE SECVENȚIALE

§ 1. Interval de încredere mărginit pentru coeficientul de regresie bazat pe o clasă de statistici de rang 389
1.1. Introducere 389
1.2. Notații și presupuneri 389
1.3. Comportarea asimptotică a unor procese empirice 392
1.4. Proprietăți asimptotice ale intervalului $I_{N(d)}$ 394
§ 2. Teste secvențiale de rang pentru regresie 395
2.1. Introducere 395
2.2. T.S.R. propuse 395
2.3. Terminarea T.S.R. cu probabilitatea unu 397
2.4. Funcția C.O. a T.S.R. 398
2.5. V.M.V.S. a T.S.R. 398
2.6. Rezultate privind eficiența asimptotică relativă 399
§ 3. Teste neparametrice pentru regresia multiplă în censorizarea progresivă 400
3.1. Introducere 400
3.2. Noțiuni preliminarii 400
3.3. Teste în S.C.P. 404
3.4. Repartiția asimptotică în ipoteza nulă 406
3.5. Repartiția asimptotică în ipoteze nenule 408
3.6. Eficiența asimptotică relativă în sensul lui Bahadur pentru teste în S.C.P. 410
Bibliografie 415

Scanned with OKEN Scanner

CONTENTS

Capter XXVIII

CORRELATION COEFFICIENTS

§ 1. Bivariate normal distribution 19
§ 2. Thesample correlation coefficient. Distribution of R in normal populations 28
§ 3. Distribution of R in non — normal populations 33
§ 4. Approximations and tables 34
§ 5. Sample covariance 41
§ 6. Circular serial correlation 43
§ 7. Noncircular serial correlation 46
§ 8. Leipnik distribution 51
§ 9. Multiple correlation coefficient 54
§ 10. Trivariate normal distributions. Partial correlation 61
§ 11. Dichotomized variables 69
11.1 Tetrachoric correlation 69
11.2. Biserial correlation 72
§ 12. Truncated bivariate and trivariate normal distributions 74
§ 13. Canonical correlations 77
§ 13.1. Normal populations 77
13.2. Nonnormal populations 79
§ 14. The correlation coefficient for directional data 82
§ 15. Ranks correlation 89
15.1. Spearman's coefficient 89
15.2. Kendall's coefficient 93
§ 16. Other measures of dependence 94
16.1. Extinction of notion of correlation coefficient for every random variables 94
16.2. A global correlation coefficient 94
16.3. The informational correlation and informational correlation coefficient 96
16.3.1. Discrete distributions 96
16.3.2. Distribution with continuous density 98
16.4. Entropic correlation and entropic correlation coefficient 100

Chapter XXIX

INFERENCES CONCERNING CORRELATION COEFFICIENTS AND COVARIANCE MATRICES

§ 1. Estimation of bivariate normal distribution parameters 105
1.1. Total sample (Unbroken sample) 105
1.2. Estimation of the correlation coefficient from a broken random sample. 110

Scanned with OKEN Scanner

1.2.1. Estimation of the unknown parameters 110
1.2.2. The integrated likelihood function 112
1.2.3. Fisher information 114
§ 2. Inferences concerning correlation coefficient with Fisher's aproximations 115
2.1. Test for a given ρ 115
2.2. Test for equality of two correlation coefficients 116
2.3. Test for homogeneity of a set of correlation coefficients 117
§ 3. Testing for independence in a bivariate normal distribution . . . 117
§ 4. Testing for independence for directional data 119
§ 5. Tests on multiple correlation coefficient and partial multiple correlation coefficients 120
5.1. Introduction 120
5.2. Maximal invariants and likelihood — ratio tests 121
5.3. Distribution of Z_2 and Z_3 123
5.4. Tests for problems i) — iv) 125
§ 6. Tests in canonical correlation analysis for non normal populations 126
§ 7. Tests about covariance matrices for nonnormal populations . . . 130
7.1. Testing the hypothesis concerning the equality between covariance matrix and a given matrix 130
7.2. Tests for sphericity 132
7.3. Test for equality of two covariance matrices 132
7.4. Testing the hypotesis concerning the equality between covariance matrix Σ and a block diagonal matrix 133

Chapter XXX

CONDITIONAL EXPECTATION. REGRESSION

§ 1. The regression curve 135
§ 2. Linear regression. The regression coefficients 136
§ 3. The correlation coefficient 138
§ 4. The correlation ratio 140
§ 5. Conection between the regression curve and cumulants 141
§ 6. Linear partial regressions 149
§ 7. Sample regression coefficients 153
§ 8. Sample coefficients 156
§ 9. Testing correlation ratios and liniarity of regression 158
§10. Applications of unbiasedness to regression 160
§11. Regression models for directional data 162

Chapter XXXI

LINEAR REGRESSION. THE FULL RANK MODEL

§ 1. Definition of general linear model. Assumptions 164
§ 2. Estimation of parameters 165
2.1. The method of least squares (L.S) 165
2.2. Properties of least squares estimator of β 169
2.3. Sums of squares. Estimations of σ^2 171
2.4. Distribution of the estimator of β. Distributions of sums of squares 174
2.5. Pure error 177

Scanned with OKEN Scanner

2.6. Confidence intervals 178
2.7. Testing hypotheses 179
§ 3. The general linear hypothesis 181
3.1. Testing general linear hypothesis 181
3.2. Estimation under the null hypothesis 183
3.3. Applications 184
§ 4. Reduced models 187
§ 5. Examining residuals 192

Chapter XXXII

MODELS NOT OF FULL RANK

§ 1. The model. Assumptions 195
§ 2. Estimation of parameters 195
2.1. The normal equations 195
2.2. Properties of normal equations solution 196
2.3. Sums of squares. Estimation of σ^2 197
2.4. Distribution of normal equations solution. Distributions of sums of squares 199
2.5. Testing hypotheses 202
§ 3. Estimable functions 203
3.1. Definition of estimable function 203
3.2. Properties of estimable function 204
3.3. Confidence intervals 206
3.4. Functions why are estimable 206
§ 4. The general linear hypotheses 209
4.1. Testable hypotheses 209
4.2. Non — testable hypotheses 214
§ 5. Restricted models 218
5.1. Introduction 218
5.2. Restrictions involving estimable functions 219
5.3. Restrictions involving non-estimable functions 220
Problems 221

Chapter XXXIII

LATENT ROOT REGRESSION ANALYSIS

§ 1. Introduction 223
§ 2. Obtaining prediction equations 224
§ 3. A geometrical interpretation of the latent roots and latent vectors of A'A 226
§ 4. Least squares and modified least squares estimators 227
§ 5. Elimination of independent variables 229
§ 6. The asymptotic properties 231

Chapter XXXIV

RIDGE REGRESSION. ESTIMATORS BY GENERALIZED INVERSE

§ 1. Ridge regression 235
1.1. Introduction 235
1.2. Ridge regression 236

Scanned with OKEN Scanner

1.3. The ridge trace 238
1.3.1. Definition of the ridge trace. Characterisation of the ridge trace 238
1.3.2. Likelihood characterization of the ridge trace 239
2.4. Mean square error properties of ridge regression 240
1.5. A general form of ridge regression 240
§ 2. The estimator and its generalizations 243
2.1. Generalized inverses 243
2.2. Properties of generalized inverse solution 245
§ 3. A class of shrinkage estimators 248
3.1. Stating the problem 248
3.2. The ridge estimator and its generalizations 250
3.3. Restricted least squares 252
§ 4. Good and optimal ridge estimators 254
4.1. Introduction 254
4.2. Good ridge factors 256
4.3. Optimal ridge factor 258
5.. Cross-validation for choosing a good ridge parameter 260

Chapter XXXV

ON FIXED—WIDTH CONFIDENCE BOUNDS FOR REGRESSION PARAMETERS

§ 1. Problem 1 266
1.1. Solution 266
1.2. Asymptotic theory for large n 271
1.3. Asymptotic properties of class $\mathscr{C}$ 272
§ 2. Problem 2 275
2.1. Solution 275

Chapter XXXVI

ROBUST REGRESSION. ESTIMATION

§ 1. Introduction 279
§ 2. M-estimates 285
2.1. The least squares when the errors are non-normal 285
2.2. M — estimates of regression 287
2.3. The bias 293
2.4. The covariance terms 294
2.5. Estimation of the covariance matrix of $\hat{\theta}$ 296
2.6. Computation of such estimates 299
2.7. Asymptotic behaviour of M — estimates for the liniar model . . . 302
§ 3. R — estimates of regression coefficients 306
3.1. The construction of an estimate of regression parameter vector based on suitable rank statistics 306
3.1.1. Notations. Assumptions 306
3.1.2. Estimation of parameters $\Delta_1, \ldots, \Delta_K$ 309
3.1.3. Estimation of parameter α 310
3.2. Estimating regression coefficients by minimizing the dispersion of the residuals 311
3.2.1. Definition of some dispersion measures and properties of the regression estimates 311

Scanned with OKEN Scanner

3.2.2. Asymptotic equivalence to Jurečková's estimates 313
3.3. Estimating regression coefficients by minimizing the dispersion function defined by a weighted Gini's mean diference 317
3.3.1. The definition and properties of the dispersion function 317
3.3.2. A general class of independent variables 320
3.3.3. Asymptotic linearity 323
3.3.4. Distribution of $\hat{\beta}$ 327
§ 4. Asymptotic relations of M — estimates and R-estimates in linear regression model 332
4.1. Assumptions and notation 332
4.2. Asymptotic distribution of $\hat{\Delta}_M - \hat{\Delta}_R$ 334
4.3. Asymptotic behaviour of $M_j(\mathbf{X}_N, \Delta)$ 335
§ 5. On estimating the density function of error variables in regression 335
5.1. Introduction 335
5.2. An estimate on the error density 337
5.3. The uniform (rectangular) case of weighted function 339
5.4. An extension to multiple regression 341
5.5. Almost sure convergence 342

Chapter XXXVII

ROBUST REGRESSION. TESTING HYPOTHESES

§1. Rank tests for the parameters of several regression lines 344
1.1. Rank tests based on the individual ranks of k different samples for the paralelism of several regression lines 344
1.1.1. Notations. Assumptions 344
1.1.2. Asymptotically optimum parametric tests 346
1.1.3. Asymptotically most powerful rank order tests 348
1 1.4. Test for a more general model 350
1 1.5. Monotonicity of regression rank statistics 353
1.2. Rank order tests for the identity of the intercepts of several regression lines 353
1.2.1. Notations. Assumptions 353
1.2.2. Optimal parametric tests 356
1.2.3. Aligned rank order tests 357
1.2.4. Asymptotic relative efficiency 358
1.3. Rank tests based on the simultaneous ranking of all observations for the identity of several regression lines 358
1.3.1. Notations and preliminaries 358
1.3.2. Asymptotic distribution of $\hat{L}_n$ 360
1.3.3. Asymptotic efficiency 361
§ 2. Rank tests of sub-hypotheses in the general linear regression 362
2.1. Signed — rank test statistics 362
2.2. Rank test statistics 365
2.3. Asymptotic relative efficiency 367
§ 3. Estimation of location parameter after a preliminary test on regression 369
3.1. Univariate case 369
3.1.1. Introduction 369
3.1.2. Estimation of location 369
3.1.3. Asymptotie distribution of the estimator θ_n^* 372

3.2. Multivariate case 375
3.2.1. Introduction 375
3.2.2. Preliminary notions and the estimations 375
3.2.3. Asymptotic distribution theory of the estimators 377
3.2.4. Asymptotic bias and dispersion matrix of the estimators 379
3.2.5. Comparison of the estimators 380
§ 4. On strong convergence of regression rank statistics 381
4.1. Introduction 381
4.2. Signed rank regression statistics 385

Chzpter XXXVIII

ROBUST REGRESSION. SEQUENTIAL PROCEDURES

§ 1. On bounded lenght confidence interval for the regression coefficient based on a class of rank statistics 389
1.1. Introduction 389
1.2. Notations and assumptions 389
1.3. Asymptotic behavior of some empirical processes 392
1.4. Asymptotic properties of confidence interval $I_{N(d)}$ 394
§ 2. Sequential rank tests for regression 395
2.1. Introduction 395
2.2. The proposed SROT 395
2.3. Termination of SROT with probability one 397
2.4. The OC function of the SROT 398
2.5. The ASN of the SROT 398
2.6. ARE results 399
§ 3. Nonparametric tests for multiple regression under progressive censoring 400
3.1. Introduction 400
3.2. Preliminary notations 400
3.3. PCS rank order test for regression 404
3.4. Asymptotic null distribution theory 406
3.5. Asymptotic nonnull distribution theory 408
3.6. Bahadur ARE of PCS tests 410
Bibliography 415

Scanned with OKEN Scanner

Capitolul XXVIII

COEFICIENȚI DE CORELAȚIE

§ 1. Repartiția normală bidimensională

Definiția 1. Vom spune că vectorul aleator (X, Y) are o repartiție normală bidimensională dacă are densitatea de repartiție:

$$p(x, y) = \frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}} \times$$

(1)

$$\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\,\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

$$-\infty < x,\ y < \infty,\ -\infty < \mu_1,\ \mu_2 < \infty,\ \sigma_1,\ \sigma_2 > 0,\ -1 < \rho < 1.$$

Definiția 2. (1) cu $\sigma_1 = \sigma_2$ și $\rho = 0$ poartă numele de densitatea de repartiție normală circulară.

(1) cu $\sigma_1 \neq \sigma_2$ însă $\rho = 0$ poartă numele de densitate de repartiție normală eliptică.

(1) cu $\mu_1 = \mu_2 = 0$ și $\sigma_1 = \sigma_2 = 1$ poartă numele de densitatea de repartiție normală bidimensională standardizată.

Din teorema III. 21 urmează că probabilitatea ca vectorul aleator (X, Y) de densitatea (1) să ia valori în elipsa

$$(2)\quad \left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x-\mu_1}{\sigma_1}\cdot\frac{y-\mu_2}{\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2 = -2(1-\rho^2)\log(1-\alpha)$$

este α.

Axa principală a acestei elipse face cu axa x-lor unghiul

$$\theta = \operatorname{arctg}\left[2\rho\,\sigma_1\sigma_2/(\sigma_1^2 - \sigma_2^2)\right].$$

Dacă $\sigma_1 = \sigma_2$ și $\rho > 0$ indiferent de valoarea numerică a lui ρ, atunci $\theta = 45°$.

Dacă $\rho = 0$ și $\sigma_1 = \sigma_2$ (adică variabilele sînt independente și au dispersii egale), atunci (2) este ecuația unui cerc.

În figura 1 se prezintă cîteva elipse.

Are loc

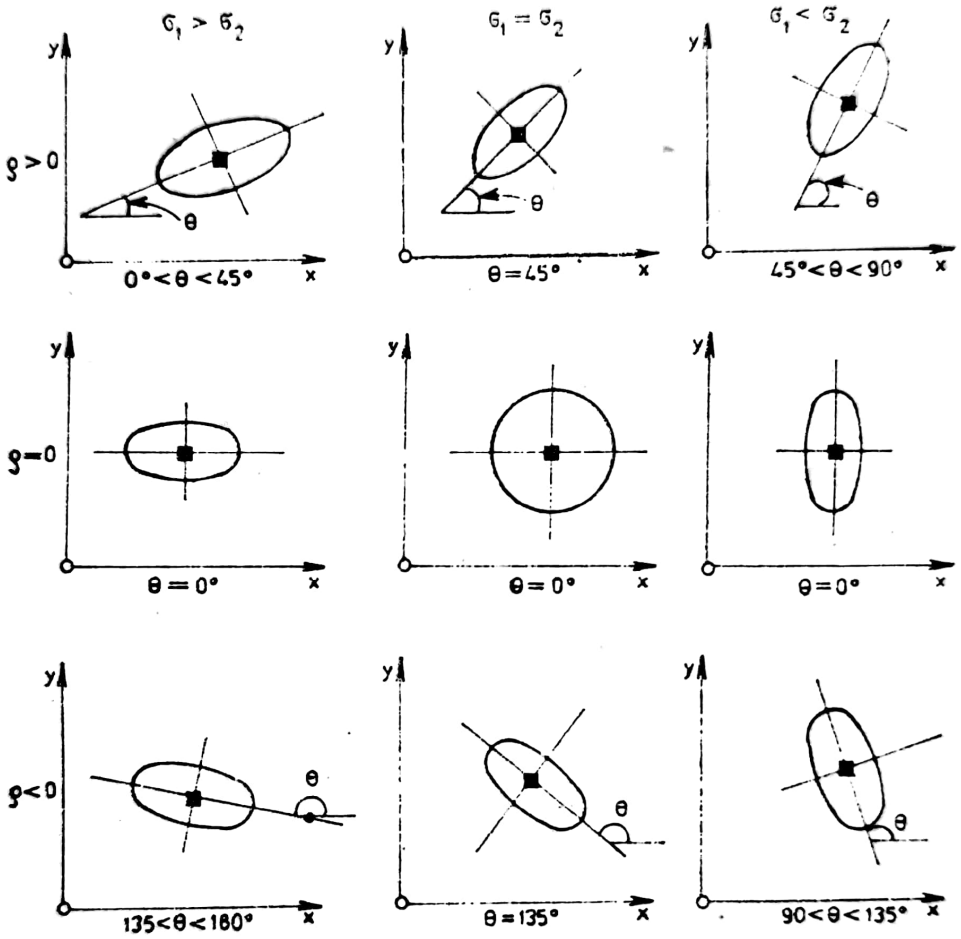
Scanned with OKEN Scanner

PROPOZIȚIA 1. *Repartițiile marginale pentru variabilele X și Y componente ale vectorului (X, Y) de densitate* (1) *sînt*

$$f_X(x) = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right] \tag{3}$$

și respectiv

$$f_Y(y) = \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]. \tag{4}$$



Punctul ● are coordonatele μ_1, μ_2 și poartă numele de centru

Fig. 1

Observație. Deoarece densitățile marginale (3) și (4) sînt densitățile corespunzătoare unor variabile normale unidimensionale,

$$M(X) = \mu_1,\ M(Y) = \mu_2,\ D^2(X) = \sigma_1^2,\ D^2(Y) = \sigma_2^2.$$

PROPOZIȚIA 2. *Covarianța variabilelor X și Y, componente ale vectorului (X, Y) de densitate* (1) *este*

$$\operatorname{cov}(X, Y) = \rho\sigma_1\sigma_2. \tag{5}$$

Scanned with OKEN Scanner

Observație. Deoarece $\sigma_1,\ \sigma_2 > 0$, covarianța este zero dacă și numai dacă $\rho = 0$. În acest caz (1) devine

$$p(x,\ y) = \frac{1}{2\pi\sigma_1\sigma_2}\exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)+\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right\} =$$

$$= \frac{1}{\sigma_1\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right\}\frac{1}{\sigma_2\sqrt{2\pi}}\left\{\exp-\frac{1}{2}\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right\} =$$

$$= f_X(x)f_Y(y)$$

ceea ce ne arată că X și Y sînt variabilele aleatoare independente. În acest caz particular (repartiția normală bidimensională)X și Y sînt independente dacă și numai dacă $\operatorname{cov}(X,\ Y) = 0$, fapt care în general nu este adevărat.

PROPOZIȚIA 3. *Densitățile de repartiție ale variabilelor X și Y condiționate de Y, respectiv X sînt date de*

$$(6)\quad f(x\,|\,y) = \frac{1}{\sqrt{2\pi}\,\sigma_1\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2\sigma_1^2(1-\rho^2)}\left(\left[x-\left(\mu_1+\rho\frac{\sigma_1}{\sigma_2}\left(y-\mu_2\right)\right)\right]^2\right\}$$

și respectiv de

$$(7)\quad f(y\,|\,x) = \frac{1}{\sqrt{2\pi}\,\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2\sigma_2^2(1-\rho^2)}\left(\left[y-\left(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)\right)\right]^2\right\}$$

Observație. Avem de-a face cu repartiții normale cu

$$(8)\quad M(X\,|\,y) = \mu_1+\rho\frac{\sigma_1}{\sigma_2}(y-\mu_2),\quad D^2(X\,|\,y) = \sigma_1^2(1-\rho^2)$$

$$(9)\quad M(Y\,|\,x) = \mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1),\quad D^2(Y\,|\,x) = \sigma_2^2(1-\rho^2).$$

PROPOZIȚIA 4. *Pentru repartiția normală bidimensională standardizată, avem :*

$$(10)\quad \begin{aligned} &\mu_{21} = \mu_{12} = 0;\\ &\mu_{31} = \mu_{13} = 3\rho;\ \mu_{22} = 1+2\rho^2;\\ &\mu_{41} = \mu_{14} = 0;\ \mu_{32} = \mu_{23} = 0;\\ &\mu_{51} = \mu_{15} = 15\rho;\ \mu_{42} = \mu_{24} = 3(1+4\rho^2);\qquad \mu_{33} = 3\rho(3+2\rho^2) \end{aligned}$$

Scanned with OKEN Scanner

și în general

$$\mu_{2r,2s} = \frac{(2r)!\,(2s)!}{2^{r+s}} \sum_{j=0}^{t} \frac{(2\rho)^{2j}}{(r-j)!\,(s-j)!\,(2j)!} \tag{11}$$

$$\mu_{2r+1,2s+1} = \frac{(2r+1)!\,(2s+1)!}{2^{r+s}} \sum_{j=0}^{t} \frac{(2\rho)^{2j}}{(r-j)!\,(s-j)!\,(2j+1)!}, \tag{12}$$

$t = \min\,(r,\ s)$

$$\mu_{r,s} = 0 \quad \text{dacă} \quad r+s = nr.\ \text{impar} \tag{13}$$

PROPOZIȚIA 4'. *Pentru repartiția normală bidimensională cu probabilitatea elementară*

$$dF = \frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_1^2} - 2\rho\frac{xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right)\right\} dx\,dy,$$

avem

$$\begin{cases} \mu_{20} = \sigma_1^2,\ \mu_{11} = \rho\sigma_1\sigma_2,\ \mu_{02} = \sigma_2^2 \\ \mu_{21} = \mu_{12} = 0;\ \mu_{30} = \mu_{03} = 0 \\ \mu_{40} = 3\sigma_1^4,\ \mu_{04} = 3\sigma_2^4 \\ \mu_{31} = 3\rho\sigma_1^3\sigma_2,\ \mu_{13} = 3\rho\sigma_1\sigma_2^3,\ \mu_{22} = (1+2\rho^2)\sigma_1^2\sigma_2^2 \end{cases} \tag{10'}$$

Demonstrație. Dacă în funcția generatoare de momente

$$G(t_1,\ t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{xt_1+yt_2}\,dF$$

facem transformarea

$$\xi = x - \sigma_1^2 t_1 - \rho\sigma_1\sigma_2 t_2$$

$$\eta = y - \rho\sigma_1\sigma_2 t_1 - \sigma_2^2 t_2$$

găsim

$$G_1(t_1,\ t_2) = \exp\left\{\frac{1}{2}(t_1^2\sigma_1^2 + 2t_1t_2\sigma_1\sigma_2\rho + t_2^2\sigma_2^2)\right\}.$$

Ținînd seamă că μ_{rs} este coeficientul lui $\frac{t_1^r}{r!}\frac{t_2^s}{s!}$ din $G(t_1,\ t_2)$, propoziția este demonstrată.

PROPOZIȚIA 5. (Kendall, Stuart (1963)). *Are loc următoarea relație de recurență:*

$$\mu_{rs} = (r+s-1)\rho\mu_{r-1,s-1} + (r-1)(s-1)(1-\rho^2)\mu_{r-2,s-2} \tag{14}$$

Scanned with OKEN Scanner

Pearson, Young (1918) tabelează μ_{rs} cu 9 zecimale pentru $r,\ s \leqslant 10$, $\rho = 0{,}00(0{,}05)1{,}00$.

Propoziția 6. *Funcția generatoare de momente comună variabilelor X și Y este*

$$M(e^{t_1X+t_2Y}) = \exp\left[-\frac{1}{2}(t_1^2 + 2\rho t_1 t_2 + t_2^2)\right] \tag{15}$$

Propoziția 7. (Kamat (1953), Nabeya (1951)). *Avem:*

$$\nu_{rs} = M(|X^r Y^s|) = \pi^{-\xi(r+s)/2}\Gamma\left(\frac{r+1}{2}\right)\Gamma\left(\frac{s+1}{2}\right)F\left(-\frac{1}{2}r, -\frac{1}{2}s;\ \frac{1}{2};\ \rho^2\right), \tag{16}$$

cu

$$F(\alpha,\ \beta;\ \gamma;\ z) = 1 + \frac{\alpha\beta}{1!\,\gamma}z + \frac{\alpha(\alpha+1)\beta(\beta+1)}{2!\gamma(\gamma+1)}z^2 + \ldots \tag{17}$$

Pentru valori mici ale lui r și s, funcțiile hypergeometrice și gama pot fi evaluate în termeni de funcții mai elementare. Obținem:

$$\begin{aligned}
\nu_{11} &= 2\pi^{-1}(\sqrt{1-\rho^2} + \rho \arcsin \rho);\\
\nu_{12} &= \nu_{21} = \sqrt{(2/\pi)}(1+\rho^2);\\
\nu_{13} &= \nu_{31} = 2\pi^{-1}\{\sqrt{1-\rho^2}(2+\rho^2) + 3\rho \arcsin \rho\};\\
\nu_{22} &= 1 + 2\rho^2;\\
\nu_{14} &= \nu_{41} = \sqrt{(2/\pi)}(3 + 6\rho^2 - \rho^4);\\
\nu_{23} &= \nu_{32} = 2\sqrt{(2/\pi)}(1 + 3\rho^2);\\
\nu_{15} &= \nu_{51} = 2\pi^{-1}\{\sqrt{1-\rho^2}(8+9\rho^2-2\rho^4) + 15\rho \arcsin \rho\};\\
\nu_{24} &= \nu_{42} = 3(1 + 4\rho^2)\\
\nu_{33} &= 2\pi^{-1}\{\sqrt{1-\rho^2}(4 + 11\rho^2) + 3\rho(3 + 2\rho^2) \arcsin \rho\}
\end{aligned} \tag{18}$$

Nabeya dă valorile lui ν_{rs} pentru $r + s \leqslant 12$. Momentele incomplete

$$\left([r,\ s] = \int_0^\infty\int_0^\infty x^r y^s p(x,\ y)\,dx\,dy\right) \tag{19}$$

au fost evaluate de Kamat (1958).

Propoziția 8. *Avem:*

$$\begin{aligned}
[r,\ s] = 2^{\frac{1}{2}(r+s)-2}\pi^{-1}\Big\{&\Gamma\left(\frac{r+1}{2}\right)\Gamma\left(\frac{s+1}{2}\right)F\left(-\frac{1}{2}r,\ -\frac{1}{2}s;\ \frac{1}{2};\ \rho^2\right) +\\
&+ 2\rho\Gamma\left(\frac{r}{2}+1\right)\Gamma\left(\frac{s}{2}+1\right)F\left(-\frac{1}{2}(r-1), -\frac{1}{2}(s-1);\ \frac{3}{2};\ \rho\right)
\end{aligned} \tag{20}$$

Scanned with OKEN Scanner

PROPOZIȚIA 9. *Avem*

$$(21)\quad [0,\ 0]=P[(X>0)\cap(Y>0)]=\frac{1}{4}+\frac{1}{2\pi}\arcsin\rho \quad \text{(Sheppard (1900))}$$

$$(22)\quad \begin{cases} [1,\ 0]=\dfrac{1}{4}\sqrt{(2/\pi)}\,(1+\rho) \\ [2,\ 0]=\dfrac{1}{4}+\dfrac{1}{2}\,\pi^{-1}[\arcsin\rho+\rho\sqrt{1-\rho^2}]; \\ [1,\ 1]=\dfrac{1}{2}\,\pi^{-1}\left[\rho\left(\dfrac{1}{2}\,\pi+\arcsin\rho\right)+\sqrt{1-\rho^2}\right]; \\ [3,\ 0]=\dfrac{1}{4}\sqrt{(2/\pi)}\,(1+\rho)^2(2-\rho)\,;\ [2,\ 1]=\dfrac{1}{4}\sqrt{(2/\pi)}\,(1+\rho)^2\,; \\ [4,\ 0]=\dfrac{1}{2}\,\pi^{-1}\left[3\left(\dfrac{1}{2}\,\pi+\arcsin\rho\right)+(5\rho-2\rho^3)\sqrt{1-\rho^2}\right]; \\ [3,\ 1]=\dfrac{1}{2}\,\pi^{-1}\left[3\rho\left(\dfrac{1}{2}\,\pi+\arcsin\rho\right)+(2+\rho^2)\sqrt{1-\rho^2}\right]; \\ [2,\ 2]=\dfrac{1}{2}\,\pi^{-1}\left[(1+2\rho^2)\left(\dfrac{1}{2}\,\pi+\arcsin\rho\right)+3\rho\sqrt{1-\rho^2}\right]. \end{cases}$$

Kamat dă tabele pentru $[r,\ s]$ cu 6 zecimale pentru $r+s\leqslant 4$; $\rho=-$ $-0{,}9(0{,}1)\ 1{,}0$.

Funcția de repartiție corespunzătoare densității normale bidimensionale standardizate este

$$(23)\quad \Phi(h,\ k;\ \rho)=(2\pi\sqrt{1-\rho^2})^{-1}\times$$
$$\times\int_{-\infty}^{h}\int_{-\infty}^{k}\exp\left[-\frac{1}{2(1-\rho^2)}(x^2-2\rho xy+y^2)\right]dx\,dy.$$

Obișnuit este tabelată cantitatea

$$(24)\quad L(h,\ k;\ \rho)=(2\pi\sqrt{1-\rho^2})^{-1}\times$$
$$\times\int_{h}^{\infty}\int_{k}^{\infty}\exp\left[-\frac{1}{2(1-\rho^2)}(x^2+2\rho xy+y^2)\right]dx\,dy.$$

PROPOZIȚIA 10. *Funcțiile* $\Phi(\cdot)$ *și* $L(\cdot)$ *sînt legate prin ecuația*

$$(25)\quad \Phi(h,\ k;\ \rho)=1-L(h,\ -\infty;\ \rho)-L(-\infty,\ k;\ \rho)+L(h,\ k;\ \rho).$$

Scanned with OKEN Scanner

Propoziția 11. *Avem* :

$$\Phi(h, \infty; \rho) = \Phi(h); \quad \Phi(\infty, k, \rho) = \Phi(k).$$

$$L(h, k; \rho) = L(k, h; \rho);$$

$$L(-h, k; \rho) - L(h, k; -\rho) = 1 - \Phi(k)$$

$$L(-h, -k; \rho) - L(h, k; \rho) = 1 - \Phi(h) - \Phi(k)$$

$$L(h, k; 0) = (1 - \Phi(h))(1 - \Phi(k))$$

$$L(h, k; 1) = \Phi(\max(h, k)).$$

$$L(h, k; -1) = \begin{cases} 0 & , \quad h + k \geqslant 0 \\ 1 - \Phi(h) - \Phi(k), & h + k \leqslant 0. \end{cases}$$

Propoziția 12. *Avem*

$$L(h, k; \rho) = \frac{1}{2\pi} \int_h^\infty \varphi(x) \int_{(k-\rho x)/\sqrt{1-\rho^2}}^\infty \varphi(y)\, dx\, dy, \tag{26}$$

unde

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} u^2\right).$$

Observație. Pentru $h > 0$, $k > 0$, $\rho > 0$, (26) reprezintă integrala din densitatea normală circulară

$$(2\pi)^{-1} \exp\left[-\frac{1}{2}(x^2 + y^2)\right]$$

pe regiunea haşurată din figura 2.

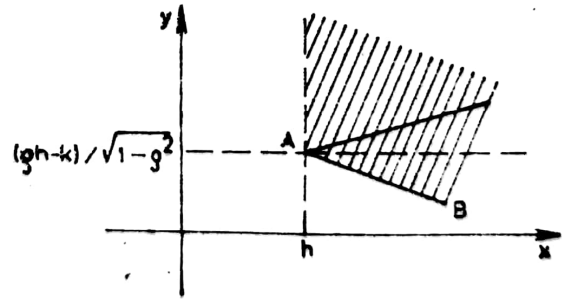


Fig. 2

Panta dreptei AB este $-\rho/\sqrt{1-\rho^2}$; unghiul dintre AB și axa y-lor este

$$\operatorname{arcctg}\left(\frac{-\rho}{\sqrt{1-\rho^2}}\right) = \frac{\pi}{2} + \arcsin \rho.$$

Scanned with OKEN Scanner

Dacă $h = k = 0$, A coincide cu 0 şi din simetria repartiţiei normale circulare

$$(27)\quad L(0,\ 0;\ \rho) = \frac{1}{2\pi}\left(\frac{\pi}{2} + \arcsin\ \rho\right) = \frac{1}{4} + \frac{1}{2\pi}\arcsin\rho \text{ (Sheeppard 1899)}$$

Sheppard (1900) dă valorile funcţiei $L(U_{0,75}, U_{0,75}; \rho)$ pentru $\pi^{-1}\arccos\rho = 0{,}00(0{,}01)0{,}80$. ($U_{0,75} = 0{,}67749$) cu 7 zecimale.

Karl Pearson (1931) dă tabele pentru $L(h, k;\ \rho)$.

Propoziţia 13. (Pólya (1949)). *Avem*:

$$(28)\quad 1 - \Phi(h) - \frac{1-\rho^2}{\rho h - k}\varphi(k)\left\{1 - \Phi\left(\frac{h-\rho k}{\sqrt{1-\rho^2}}\right)\right\} < L(h,\ k;\ \rho) < 1 - \Phi(h)$$

pentru $0 < \rho < 1;\ \rho h - k > 0$.

Observaţie. Deoarece $\rho h - k > 0$, urmează că $h - \rho k > 0$ (dacă $h > 0$) şi astfel (28) este de forma

$$(28')\quad 1 - \Phi(h) - \Delta < L(h,\ k;\rho) < 1 - \Phi(h)$$

cu

$$0 < \Delta < (2\sqrt{2\pi})^{-1}(1-\rho^2)(\rho h - k)^{-1}.$$

Propoziţia 14 (Zelen, Severo (1960)). *Avem*

$$(29)\quad L(h,\ k;\ \rho) = L(h,\ 0;\ \rho(h,\ k)) + L(k,\ 0;\ \rho(k,\ h)) - \frac{1}{2}(1 - \delta_{hk}),$$

unde:

$$\rho(h,\ k) = \frac{(\rho h - k)f(h)}{\sqrt{h^2 - 2\rho hk + k^2}};$$

$$f(h) = \begin{cases} 1, & h > 0 \\ -1, & h < 0 \end{cases}$$

$$\delta_{hk} = \begin{cases} 0\ , & Sg(h)Sg(k) = 1 \\ 1\ , & \text{în rest} \end{cases}$$

cu

$$Sg(h) = 1,\ h \geqslant 0$$

$$Sg(h) = -1,\ h < 0.$$

Observaţie. Propoziţia 14 permite evaluarea funcţiei $L(h,\ k;\ \rho)$ din tabele cu $k = 0$.

Propoziţia 15 (Sheppard (1900)). *Avem*:

$$L(h,\ k;\ \rho) = V\left(h,\ \frac{k-\rho h}{\sqrt{1-\rho^2}}\right) + V\left(k,\ \frac{h-\rho k}{\sqrt{1-\rho^2}}\right) +$$

$$(30)\quad +1 - \frac{1}{2}(\Phi(h) + \Phi(k)) - \frac{\arccos\rho}{2\pi}$$

Scanned with OKEN Scanner

unde

$$V(h,\ k) = \int_0^h \varphi(x) \int_0^{kx/h} \varphi(y)\, dx\, dy. \tag{31}$$

Observație. $V(h,\ k)$ pentru $h > 0,\ k > 0$ este integrala din densitatea normală circulară pe triunghiul OPH ($P(h,\ k)$, $H(h,\ 0)$ (fig. 3). Similar,

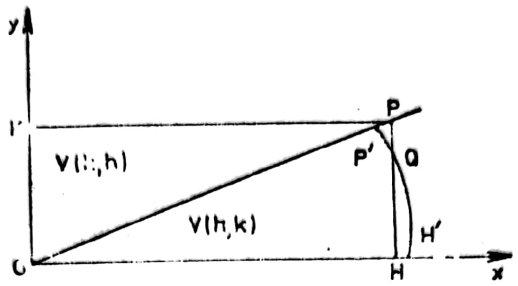


Fig. 3

$V(k,\ h)$ este integrala pe triunghiul OKP. Deoarece integrala pe dreptunghiul $OHPK$ este

$$P[(0 < X < h) \cap (0 < Y < k)],$$

avem:

$$V(h,\ k) + V(k,\ h) = \left(\Phi(h) - \frac{1}{2}\right)\left(\Phi(k) - \frac{1}{2}\right). \tag{32}$$

Pentru celelalte valori ale lui h și k, $V(h,\ k)$ este definită de

$$V(h,\ -k) = -V(h,\ k) = V(-h,\ k) \tag{33}$$

și prin urmare $V(-h,\ -k) = V(h,\ k)$.

Nicholson (1943) tabelează $V(h,\ k)$ cu 6 zecimale pentru $h,\ k = 0{,}0(0{,}1)3{,}1$; $\infty(V(h,\ \infty) = \Phi(h) - \frac{1}{2}$; $V(\infty,\ k) = 0)$.

Nicholson folosește formula:

$$V(h,\ k) = (2\pi)^{-1}\left[\lambda(1 - e^{-m}) - \frac{1}{3}\lambda^3(1 - e^{-m} - me^{-m}) + \right.$$

$$\left. + \frac{1}{5}\lambda^5\left(1 - e^{-m} - me^{-m} - \frac{m^2}{2!}e^{-m}\right) - \ldots\right],$$

cu $\lambda = k/h < 1$; $m = \frac{1}{2}h^2$. Pentru $\lambda > 1$, poate fi folosit (32).

Dacă în locul lui h și k se ia h și $\lambda(=k/h)$ interpolarea este ușurată

Observație. Avem

$$V(\infty,\ \lambda \cdot \infty) = \lim_{h \to \infty} V(h,\ \lambda h) = \frac{1}{2}\pi \operatorname{arctg} \lambda. \tag{35}$$

Scanned with OKEN Scanner

Pentru $h \geq 5{,}6$ și $0{,}1 \leq \lambda \leq 1$, $V(h, \lambda h)$ concordă cu $V(\infty, \lambda \cdot \infty)$ la a 7-a zecimală.

Avem

$$V(\lambda \cdot \infty, \infty) = \lim_{h \to \infty} V(\lambda h, h) = 2\pi \operatorname{arctg} \lambda \tag{36}$$

însă aproximația

$$V(\lambda h, h) \simeq \frac{1}{2\pi} \operatorname{arctg} \lambda - \frac{1}{2} \{1 - \Phi(\lambda h)\} \tag{37}$$

are loc cu o eroare mai mică decît $\frac{1}{2} \times 10^{-7}$ pentru $h \geq 5{,}6$.

Cadwell (1951) aproximează $V(h, k)$ pentru $k/h = \lambda$ mic, înlocuind PH prin arcul $P'H'$ cu centrul în O și raza aleasă astfel încît $P'QP$ și $H'QH$ să aibe arii egale (Q fiind punctul de intersecție dintre arcul $P'H'$ și PH). Aproximația rezultată este

$$V(h, \lambda h) \simeq (2\pi)^{-1} \operatorname{arctg} \lambda \left[1 - \exp\left(-\frac{\frac{1}{2} h^2 \lambda}{\operatorname{arctg} \lambda}\right)\right]. \tag{38}$$

§ 2. Coeficientul de corelație de selecție R. Repartiția statisticii R în populații normale

DEFINIȚIA 3. Coeficientul de corelație de selecție, bazat pe n perechi de valori observate a două caracteristici, reprezentate de variabilele aleatoare (X_t, Y_t), $t = 1, \ldots, n$ este definit

$$R = \frac{\sum_{t=1}^{n} (X_t - \bar{X})(Y_t - \bar{Y})}{\left[\sum_{t=1}^{n} (X_t - \bar{X})^2 \sum_{t=1}^{n} (Y_t - \bar{Y}^2\right]^{1/2}} = \frac{m_{11}}{(m_{20} m_{02})^{1/2}}. \tag{39}$$

R se mai poate scrie:

$$R = \frac{\sum X_t Y_t - (\sum X_t)(\sum Y_t)/n}{[\{X_t^2 - (\sum X_t)^2/n\}\{\sum Y_t^2 - (\sum Y_t)^2/n\}]^{1/2}} \tag{39'}$$

PROPOZIȚIA 16. $-1 \leq r \leq 1$.

În acest paragraf dăm repartiția statisticii R cînd $(X_1, Y_1), \ldots, (X_n, Y_n)$ este o selecție asupra unui vector normal bidimensional (X, Y) de densitate (1).

Au loc:

PROPOZIȚIA 17. *În repartiția normală bidimensională cu densitatea* (1) *cu* $\mu_1 = \mu_2 = 0$ *și* $\sigma_1 = \sigma_2 = 1$, *statistica* $V = R(1 - R^2)^{-1/2}$ *are pentru valori fixate ale variabilelor* $X_1, X_2, \ldots X_n$ *aceeași repartiție ca și*

$$(n-2)^{-1/2} t\left(n - 2, \sqrt{\sum_{t=1}^{n} (X_t - \bar{X})^2} \frac{\rho}{\sqrt{1 - \rho^2}}\right).$$

Scanned with OKEN Scanner

Pentru a obține repartiția necondiționată a statisticii V, trebuie să calculăm valoarea medie a densității astfel obținute cu repartiția variabilelor $X_1, X_2, \ldots, X_n$. Deoarece densitatea de repartiție depinde de $X_1, \ldots, X_n$ numai prin funcția $S = \sum_1^n (X_i - \bar{X})^2$ folosim faptul că $S \sim \chi^2(n-1)$ (teorema III.5). Densitatea de repartiție a statisticii V condiționate de S este

$$p_V(v|S) = \frac{\exp\left[-\frac{\rho^2 S}{2(1-\rho^2)}\right]}{\sqrt{\pi}\,\Gamma\left(\frac{1}{2}n-1\right)} (1+v^2)^{-\frac{1}{2}(n-1)} \times$$

$$\times \sum_{j=0}^{\infty} \frac{\Gamma\left(\frac{n-1+j}{2}\right)}{j!} \left(\frac{2\rho^2 v^2 S}{(1-\rho^2)(1+v^2)}\right)^{j/2}. \tag{40}$$

Deoarece

$$p_S(s) = \frac{1}{2^{\frac{n-2}{2}}\Gamma\left(\frac{n-1}{2}\right)} s^{(n-3)/2} e^{-\frac{1}{2}s} \qquad (s>0)$$

și

$$\int_0^\infty s^{j/2} j \exp\left[-\frac{1}{2}\rho^2 s(1-\rho^2)^{-1}\right] p_S(s)\,ds =$$

$$= \frac{2^{j/2}\Gamma\left(\frac{n-1+j}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} (1-\rho^2)^{(n+j-1)/2},$$

$$p_V(v) = \frac{(1-\rho^2)^{(n-1)/2}}{\sqrt{\pi}\,\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{1}{2}n-1\right)} (1+v^2)^{-(n-1)/2} \times$$

$$\times \sum_{j=0}^{\infty} \frac{(2\rho)^j \left[\Gamma\left(\frac{n-1+j}{2}\right)\right]^2}{j!} \left(\frac{v^2}{1+v^2}\right)^{j/2}. \tag{41}$$

Făcînd transformarea $V = R(1-R^2)^{-1/2}$ obținem

$$p_R(r) = \frac{(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2}}{\sqrt{\pi}\,\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{1}{2}n-1\right)} \times$$

$$\times \sum_{j=0}^{\infty} \frac{\left[\Gamma\left(\frac{n-1+j}{2}\right)\right]^2}{j!} (2\rho r)^j, \quad r^2 \leqslant 1. \tag{42}$$

Scanned with OKEN Scanner

Constanta de multiplicare poate fi exprimată într-o formă alternativă folosind identitatea :

$$\sqrt{\pi}\,\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{1}{2}n-1\right)=2^{-(n-3)}\,\pi(n-3)\,!.$$

Există un număr de alte forme în care membrul drept al relaţiei (42) poate fi exprimat. Acestea includ :

$$p_R(r)=\frac{(n-2)\,(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2}}{\pi}\int_0^\infty\frac{dw}{(\mathrm{ch}w-\rho r)^{n-1}} \tag{43}$$

(teorema III.14)

$$p_R(r)=\frac{(n-2)(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2}}{\pi}\int_1^\infty\frac{dw}{(w-\rho r)^{n-1}(w^2-1)^{1/2}} \tag{44}$$

$$p_R(r)=\frac{(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2}}{\pi\Gamma(n-2)}\frac{d^{n-2}}{d(r\rho)^{n-2}}\left\{\frac{\arccos(-\rho r)}{(1-\rho^2r^2)^{1/2}}\right\} \tag{45}$$

(teorema III.14)

$$p_R(r)=\frac{(n-2)\,(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2}}{\sqrt{2}(n-1)B\left(\frac{1}{2},\,n-\frac{1}{2}\right)(1-\rho r)^{n-3/2}}\times$$

$$\times F\left(\frac{1}{2},\,\frac{1}{2};\,n-\frac{1}{2};\,\frac{1}{2}(1+\rho r)\right) \tag{46}$$

$$p_R(r)=\frac{(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2}}{\pi\Gamma(n-2)}\left(\frac{\partial}{\sin\theta\partial\theta}\right)^{n-2}\frac{\theta}{\sin\theta},\ \theta=\arccos(\rho r). \tag{47}$$

(47) a fost obţinută de Fisher (1915). Cercetări anterioare se datoresc lui Student (1908) şi Soper (1913).

Propoziţia 18. (Garwood (1933)). $F_{n,R}(r)=P(R\leqslant r)$, *n fiind volumul selecţiei, are expresia* :

$$F_{3,R}(r)=\pi^{-1}[\arccos(-r)-\sqrt{(1-r^2)/(1-\rho^2r^2)}\arccos(-\rho r)], \tag{48}$$

$$F_{4,R}(r)=\rho^{-1}\sqrt{(1-\rho^2)\,(1-r^2)}\,F_{3,R}(r)-\pi^{-1}[\rho^{-1}\sqrt{1-\rho^2}-\arccos\rho], \tag{49}$$

$$F_{5,R}(r)=\frac{1}{2}\rho^{-1}\sqrt{(1-\rho^2)(1-r^2)}F_{4,R}(r)-\frac{1}{2}r(1-r^2)F_{3,R}(r)-$$

$$-\pi^{-1}\left[\frac{1}{2}\rho^{-1}(1+\rho^2)\sqrt{\frac{1-r^2}{1-\rho^2r^2}}\arccos(-\rho r)-\arccos(-r)\right], \tag{50}$$

$$F_{6,R}(r)=\frac{1}{3}\rho^{-1}\sqrt{(1-\rho^2)(1-r^2)}\,F_{5,R}(r)+\frac{1}{3}\rho^{-2}(1-\rho^2)rF_{4,R}(r)- \tag{51}$$

$$-\frac{1}{3}\rho^{-3}(1-\rho^2)^{3/2}\sqrt{1-r^2}\,F_{3,R}(r)+\pi^{-1}\left[\frac{1}{3}\rho^{-3}(1-4\rho^2)\sqrt{1-\rho^2}-\arccos\rho\right].$$

Scanned with OKEN Scanner

PROPOZIȚIA 19. (Garwood (1933)). *Pentru* $n = 2s + 3$ *funcția de repartiție a coeficientului de corelație de selecție este dată de*

$$F_R(r) = \pi^{-1}\arccos r - (1 - r^2)^{1/2}[(2s!)]^{-1}\pi^{-1}(1 - \rho^2)^{s+1} \times$$

$$\times \left[\left\{ \Delta^s \rho^{2s} - \ldots - C_s^1 \frac{\partial^2}{\partial \rho^2} \Delta^{s-1} \rho^{2s-2} + C_s^2 \frac{\partial^4}{\partial \rho^4} \Delta^{s-2} \rho^{2s-4} + \ldots + (-1) \frac{\partial^{2s}}{\partial \rho^{2s}} \right\} \times \right. \tag{52}$$

$$\left. \times \frac{\rho}{1-\rho^2} \frac{\arccos(-\rho r)}{(1 - \rho^2 r^2)^{1/2}} \right].$$

Observație. Cînd n crește, expresiile pentru funcția de repartiție devin mai complicate.

PROPOZIȚIA 20. (Ruben (1963)) *Avem*:

$$P(R \leqslant 0) = P\left[t(n-1) \leqslant \frac{-\rho\sqrt{n-1}}{\sqrt{1-\rho^2}} \right]. \tag{53}$$

Demonstrație. Deoarece $(R \leqslant 0)$ este echivalent cu

$$\sum_{i=1}^{n} (X_i - \bar{X})\bar{Y}_i \leqslant 0$$

trebuie să evaluăm probabilitatea acestui eveniment. Ținînd seama că $\sum_{j=1}^{n} (X_j - \bar{X})^2 \sim \chi^2(n-1)$ avem (cu $U \sim N(0, 1)$)

$$\Phi\left(\frac{-\rho}{\sqrt{1-\rho^2}} \sqrt{\sum (X_j - \bar{X})^2} \right) = P\left[\frac{U}{\sqrt{\sum (X_j - \bar{X})^2}} \leqslant \frac{-\rho}{\sqrt{1-\rho^2}} \right] =$$

$$= P\left[\frac{t(n-1)}{\sqrt{n-1}} \leqslant -\frac{\rho}{\sqrt{1-\rho^2}} \right] = P\left[t(n-1) \leqslant \frac{-\rho\sqrt{n-1}}{\sqrt{1-\rho^2}} \right],$$

ceea ce demonstrează propoziția.

PROPOZIȚIA 21. (Ghosh (1966)). *Momentele statistice* R, *exprimate cu ajutorul funcțiilor hipergeometrice, sînt*

$$\mu_1' = c_n \rho F\left(\frac{1}{2}, \frac{1}{2}; \frac{1}{2}(n+1); \rho^2 \right), \tag{54}$$

$$\mu_2' = 1 - \frac{(n-2)(1-\rho^2)}{n-1} F\left(1,1; \frac{1}{2}(n+1); \rho^2 \right), \tag{55}$$

$$\mu_3' = c_n \left[\rho F\left(\frac{1}{2}, \frac{1}{2}; \frac{1}{2}(n+1); \rho^2 \right) - \rho^{-1}(n-1)(n-2) \right] \times \tag{56}$$

$$\times \left\{ F\left(\frac{1}{2}, \frac{1}{2}; \frac{1}{2}(n-1); \rho^2 \right) - F\left(\frac{1}{2}, \frac{1}{2}; \frac{1}{2}(n+1); \rho^2 \right) \right\},$$

$$\mu_4' = 1 + \frac{(n-2)(n-4)(1-\rho^2)}{2(n-1)} F\left(1,1; \frac{1}{2}(n+1); \rho^2 \right) - \tag{57}$$

$$- \frac{n(n-2)(1-\rho^2)}{4\rho^2} \left[F\left((1,1; \frac{1}{2}(n+1); \rho^2 \right) - 1 \right],$$

Scanned with OKEN Scanner

unde

$$c_n = \frac{2}{n-1}\left[\Gamma\left(\frac{n}{2}\right)/\Gamma\left(\frac{n-1}{2}\right)\right]^2.$$

PROPOZIȚIA 22. (Ghosh (1966)). *Pentru* $m = (n+6)$, *avem pentru* R

$$\mu_1' = \rho - \frac{1}{2}\rho(1-\rho^2)m^{-1}\left[1 + \frac{9}{4}(3+\rho^2)m^{-1} + \right.$$

(58)

$$\left. + \frac{3}{8}(121 + 70\rho^2 + 25\rho^4)m^{-2}\right] + \mathbf{0}(m^{-4}),$$

$$\mu_2 = \frac{(1-\rho^2)^2}{m}\left[1 + \frac{1}{2}(14+11\rho^2)m^{-1} + \frac{1}{2}(98+130\rho^2+75\rho^4)m^{-2}\right] + \mathbf{0}(m^{-4})$$

(59)

$$\mu_3 = -\frac{\rho(1-\rho^2)^3}{m^2}\left[6 + (69+88\rho^2)m^{-1} + \right.$$

(60)

$$\left. + \frac{3}{4}(797 + 1691\rho^2 + 1560\rho^4)m^{-2}\right] + \mathbf{0}(m^{-5}),$$

$$\mu_4 = \frac{3(1-\rho^2)^4}{m^2}\left[1 + (12+35\rho^2)m^{-1} + \right.$$

(61)

$$\left. + \frac{1}{4}(436 + 2028\rho^2 + 3025\rho^4)m^{-2}\right] + \mathbf{0}(m^{-5}).$$

$$\beta_1 = \frac{\rho^2}{m}[36 + 6(12 + 77\rho^2)m^{-1} -$$

(62)

$$- (162 - 1137\rho^2 - 6844\rho^4)m^{-2}] + \mathbf{0}(m^{-4})$$

(semnul lui $\sqrt{\beta_1}$ este opus semnului lui ρ).

$$\beta_2 = 3 - \frac{3}{m}\left[2(1 - 12\rho^2) + (10 + 14\rho^2 - 387\rho^4)m^{-1} + \right.$$

(63)

$$\left. + \frac{1}{2}(100 + 832\rho^2 + 1503\rho^4 - 14202\rho^6)m^{-2}\right] + \mathbf{0}(m^{-4}).$$

Scanned with OKEN Scanner

Observație. Reamintim (formula (III. 25)) că deplasarea lui R ca estimator al lui ρ este aproximativ $-\frac{1}{2n}\rho(1-\rho^2)$ și că

$$D^2(R) \simeq \frac{(1-\rho^2)^2}{n}. \tag{64}$$

Se pune întrebarea, există o funcție $g(R)$ astfel încît

$$M[g(R)] = g(\rho)$$

identic în ρ?. Hotelling (1953) a arătat că dacă g nu depinde de n, $g(R)$ poate fi numai o funcție liniară de $\arcsin R$. Un rezultat mai precis este conținut în

PROPOZIȚIA 23. (Harley (1954), 1956), Daniels, Kendall (1958). *Avem*

$$M(\arcsin R) = \arcsin \rho.$$

PROPOZIȚIA 24 (Hotelling (1953)). *Avem*:

$$r\frac{\partial p_R(r)}{\partial r} + \frac{(n-3)r^2}{1-r^2}p_R(r) = \rho\frac{\partial p_R(r)}{\partial \rho} + \frac{n\rho^2}{1-\rho^2}p_R(r). \tag{65}$$

Cu notația:

$$p_{n,R}(r) = p_R(r)$$

cînd volumul selecției este n, are loc

PROPOZIȚIA 25 (Soper, Young, Cave, Lee, Pearson (1915—1917))

$$(n-1)(n-2)\{1-(\rho R)^2\}p_{n+1,R}(r) =$$

$$= (2n-1)(n-2)\rho\sqrt{1-\rho^2}\,R\sqrt{1-R^2}\,p_{n,R}(r) +$$

$$+(n-1)^2(1-\rho^2)(1-r^2)p_{n-1,R}(r). \tag{66}$$

§ 3. Repartiția statisticii R în populații care nu sînt normale

Repartiția lui R pentru selecții care nu provin din populații normale a fost obținută numai pentru anumite cazuri particulare.

Quensel (1938) presupune că cumulanții de ordin mai mare decît 4 sînt neglijabili și că coeficientul de corelație al populației (ρ) este zero. Gayen (1951) extinde această lucrare pentru $\rho \neq 0$.

Notînd densitatea $p_R(r)$ prin $f(r, \rho)$ are loc

PROPOZIȚIA 26 (Gayen (1951)) *Avem*:

$$P_R(r) = f(r, \rho) + \frac{n-1}{8n(n-1)}\left\{L_{4,1}\frac{\partial f}{\partial \rho} + L_{4,2}\frac{\partial^2 f}{\partial \rho^2}\right\} +$$

$$+\frac{n-2}{12n(n+1)(n+3)}\left\{L_{6,1}\frac{\partial f}{\partial \rho} + L_{6,2}\frac{\partial^2 f}{\partial \rho^2} + L_{6,3}\frac{\partial^3 f}{\partial \rho^3}\right\} \tag{67}$$

Scanned with OKEN Scanner

unde

$$L_{4;1} = 3\rho(\gamma_{40} + \gamma_{04}) - 4(\gamma_{31} + \gamma_{13}) + 2\rho\,\gamma_{22},$$

$$L_{4;2} = \rho^2(\gamma_{40} + \gamma_{04}) - 4\rho(\gamma_{31} + \gamma_{13}) + 2(2 + \rho^2)\gamma_{22},$$

$$L_{6;1} = -15\rho(\gamma_{30}^2 + \gamma_{03}^2) - 9\rho\left(1 + \frac{2}{n-2}\right)(\gamma_{21}^2 + \gamma_{12}^2) -$$

$$-\frac{6}{n-2}\gamma_{30}\gamma_{03} + 6\left(2 + \frac{1}{n-2}\right)\gamma_{21}\gamma_{12} + 18(\gamma_{30}\gamma_{21} + \gamma_{03}\gamma_{12}) +$$

$$+\frac{18\rho}{n-2}(\gamma_{30}\gamma_{12} + \gamma_{03}\gamma_{21}),$$

$$L_{6;2} = -9\rho^2(\gamma_{30}^2 + \gamma_{03}^2) - 3\left(4 + 5\rho^2 - \frac{2(2-5\rho^2)}{n-2}\right)(\gamma_{21}^2 + \gamma_{12}^2) -$$

$$-\frac{18\rho}{n-2}\gamma_{30}\gamma_{03} + 18\rho\left(2 + \frac{1}{n-2}\right)\gamma_{21}\gamma_{12} + 30\rho(\gamma_{30}\gamma_{21} + \gamma_{03}\gamma_{12}) -$$

$$-6\left(2 + \frac{2-5\rho^2}{n-2}\right)(\gamma_{30}\gamma_{12} + \gamma_{03}\gamma_{21}),$$

$$L_{6;3} = -\rho^3(\gamma_{30}^2 + \gamma_{03}^2) - 3\rho\left(2 + \rho^2 - \frac{2(1-\rho^2)}{n-2}\right)(\gamma_{21}^2 + \gamma_{12}^2) +$$

$$+2\left(1 + \frac{3(1-\rho^2)}{n-2}\right)\gamma_{30}\gamma_{03} + 6\left(1 + 2\rho^2 - \frac{1-\rho^2}{n-2}\right)\gamma_{21}\gamma_{12} +$$

$$+6\rho^2(\gamma_{30}\gamma_{21} + \gamma_{03}\gamma_{12}) - 6\rho\left(1 + \frac{1-\rho^2}{n-2}\right)(\gamma_{30}\gamma_{12} + \gamma_{03}\gamma_{21}),$$

unde $\gamma_{ij} = \varkappa_{ij}\varkappa_{20}^{-i/2}\varkappa_{02}^{-j/2}$

Cook (1951 a, b) obţine expresii pentru primele 4 momente ale statisticii R în funcţie de cumulanţii repartiţiei.

Al doilea şi respectiv al treilea termen din (67) pot fi privite drept corecţii pentru exces şi oblicitate în cazul în care $f(r, \rho)$ este normală. Gayen (1951) dă aceste corecţii pentru anumite cazuri particulare. Cînd ρ este zero, termenii sînt mici chiar pentru valori ale lui n mai mici decît 4.

§ 4. Aproximaţii şi tabele

Din § 2, rezultă că repartiţia lui R este atît de complicată încît este foarte dificil s-o folosim fără aproximaţii sau tabele. Notăm cu această repartiţie este importantă deoarece este repartiţia coeficientului de corelaţie de selecţie în selecţii din populaţii normale bidimensionale şi nu pentru

că este succeptibilă de ajustare. Gayen discută repartiția lui $z' = \operatorname{arcth} R$, obținînd expresii pentru valoarea medie, dispersie, β_1 și β_2.

Tabelul 1

	R	z'
Valoarea medie	$\rho + \frac{1}{n}\left[-\frac{1}{2}\rho(1-\rho^2) + \frac{1}{8}L_{4,1}\right]$	$\frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{1}{n-1} \times$ $\times\left[\frac{1}{2}\rho + \frac{1}{8(1-\rho^2)^2}\{\rho(3-\rho^2) \times\right.$ $\times(\gamma_{40}+\gamma_{04}) - 4(1+\rho^2)(\gamma_{31}+\gamma_{13}) +$ $\left.+ 2\rho(5+\rho^2)\gamma_{22}\}\right]$
Dispersia	$\frac{1}{n}\left[(1-\rho^2)^2 + \frac{1}{4}L_{4,2}\right]$	$\frac{1}{n-1}\left[1 + \frac{1}{4(1-\rho^2)^2} \times\right.$ $\times\{\rho^2(\gamma_{40}+\gamma_{04}) - 4\rho(\gamma_{31}+\gamma_{13}) +$ $\left.+ 2(2+\rho^2)\gamma_{22}\}\right]$

David (1938) dă valorile densității $p_R(r)$ cu 4 zecimale și a funcției de repartiție $F_R(r)$ cu 4 zecimale pentru $n = 3(1)\ 25,\ 50,\ 100,\ 200,\ 400$. Pentru $n \leqslant 25$, $r = -1{,}00(0{,}05)\ 1{,}00$ pentru $\rho = 0{,}0(0{,}1)\ 0{,}4$; $r = -1{,}00\ (0{,}05)\ 0{,}600\ (0{,}025)\ 1{,}000$ pentru $\rho = 0{,}5\ (0{,}1)\ 0{,}9$ cu valori adiționale pentru $r = 0{,}80\ (0{,}01)\ 0{,}900\ (0{,}005)\ 1{,}000$ pentru $\rho = 0{,}9$.

Introducerea la aceste tabele conține observații interesante asupra repartiției lui R.

Pentru cele mai multe scopuri practice, aproximațiile repartiției lui R folosesc transformarea:

$$z' = \operatorname{arcth} R = \frac{1}{2}\log\left(\frac{1+R}{1-R}\right).$$

Această transformare stabilizează dispersia (din (64))

$$D^2(R) \simeq (1-\rho^2)^2 n^{-1}$$

și

$$\int (1-\rho^2)^{-1}\,d\rho = \frac{1}{2}\log\frac{1+\rho}{1-\rho}. \tag{68}$$

Această abordare nu a fost folosită explicit de Fisher (1915) cînd a sugerat această transformare.

Scanned with OKEN Scanner

PROPOZIȚIE 27 (Fisher (1921)). *Avem*

$$\mu_1'(z') \simeq \frac{1}{2}\log\frac{1+\rho}{1-\rho}+\frac{1}{2}\rho(n-1)^{-1}\left[1+\frac{1}{4}(5+\rho^2)(n-1)^{-1}\right]$$

(69)

$$\mu_2(z')\simeq(n-1)^{-1}\left[1+\frac{1}{2}(4-\rho^2)(n-1)^{-1}+\frac{1}{6}(22-6\rho^2-3\rho^4)(n-1)^{-2}\right]$$

(70)

(71) $$\mu_3(z') \simeq (n-1)^{-3}\rho^3$$

$$\mu_4(z')\simeq 3(n-1)^{-2}\left[1+\frac{1}{3}(14-3\rho^2)(n-1)^{-1}+\frac{1}{12}(184-48\rho^2-21\rho^4)(n-1)^{-2}\right]$$

(72)

(73) $$\beta_1(z') \simeq (n-1)^{-3}\rho^6$$

(74) $$\beta_2(z') \simeq 3+2(n-1)^{-1}+(4+2\rho^2-3\rho^4)(n-1)^{-2}.$$

Observație i) Gayen (1951) corectează aceste valori.

ii) De asemenea poate fi văzut Nabeya (1951). Comparînd (73) și (74) cu (62) și (63), observăm că valorile (β_1, β_2) pentru z' sînt mult mai apropiate de valorile normale (0,3) decît valorile lui R.

Cea mai folosită aproximație privește z' ca fiind repartizată $N\left(\frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right), (n-3)^{-1}\right)$. Valoarea $(n-3)^{-1}$ este obținută notînd că

$$(n-1)^{-1}+\frac{1}{2}(4-\rho^2)(n-1)^{-2} = n^{-1}+n^{-2}+\ldots+\frac{1}{2}(4-\rho^2)n^{-3}\ldots$$

$$=n^{-1}\left[1+\left(3-\frac{1}{2}\rho^2\right)n^{-1}\ldots\right]\simeq n^{-1}[1-3n^{-1}]^{-1}=(n-3)^{-1}.$$

Pentru a îmbunătăți aproximația, valoarea medie poate fi mărită cu $(2n-5)^{-1}\rho$.

O altă aproximație poate fi bazată pe considerarea structurii variabilei aleatoare R. În § 2 am văzut că pentru o mulțime fixată de valori $X_1, X_2, \ldots, X_n$, $R(1-R^2)^{-1/2}$ este repartizată ca

$$(n-2)^{-1/2}t'\left(n-2, \sqrt{\sum_{j=1}^{n}(X_j-\bar{X})^2}\rho/\sqrt{1-\rho^2}\right),$$

adică

$$\left(U+\sqrt{\sum_{j=1}^{n}(X_j-\bar{X})^2}\rho/\sqrt{1-\rho^2}\right)/\chi(n-2)$$

unde $U \sim N(0,1)$ și U și $\chi^2(n-2)$ sînt independente. Luînd media în raport cu repartiția comună a X-lor, obținem că $R(1-R^2)^{-1/2}$ este repartizată ca

$$[U + \chi(n-1)\rho(1-\rho^2)^{-1/2}]/\chi(n-2), \tag{75}$$

unde U, $\chi(n-1)$ și $\chi(n-2)$ sînt independente. Această reprezentare a fost obținută de Ruben ((1963), (1966)) și este folosită ca bază pentru următoarele aproximații.

Din reprezentarea (75) urmează că

$$P(R \leqslant r) = P[U + \rho(1-\rho^2)^{-1/2}\chi(n-2) - r(1-r^2)^{-1/2}\chi(n-1) \leqslant 0]. \tag{76}$$

Dacă n nu este foarte mic, $\chi(n-2)$ și $\chi(n-1)$ pot fi aproximate prin variabile normale folosind aproximația lui Fisher conform căreia $\{\sqrt{2\chi^2(n)} - \sqrt{2n-1}\}$, este aproximativ, repartizată $N(0,1)$ (teorema I.4). Atunci

$$U + \rho(1-\rho^2)^{-1/2}\chi(n-2) - r(1-r^2)^{-1/2}\chi(n-1)$$

este aproximativ repartizată normal cu valoarea medie

$$\rho(1-\rho^2)^{-1/2}\left(n-\frac{5}{2}\right)^{1/2} - r(1-r^2)^{-1/2}\left(n-\frac{3}{2}\right)^{1/2}$$

cu abaterea medie pătratică

$$\left[1 + \frac{1}{2}\rho^2(1-\rho^2)^{-1} + \frac{1}{2}r^2(1-r^2)^{-1}\right]^{1/2}.$$

Deci, din (76), avem

$$P(R \leqslant r) \simeq \Phi\left(\frac{r(1-r^2)^{-1/2}\left(n-\frac{3}{2}\right)^{1/2} - \rho(1-\rho^2)^{-1/2}\left(n-\frac{5}{2}\right)^{1/2}}{\left[1 + \frac{1}{2}r^2(1-r^2)^{-1} + \frac{1}{2}\rho^2(1-\rho^2)^{-1}\right]^{1/2}}\right). \tag{77}$$

Stammbergèr (1968) a construit o nomogramă din care este posibil să se determine una dintre cantitățile $P(R \leqslant r)$, r sau n, date celelalte două valori.

Cu notațiile

$$u = \frac{r-\rho}{1-\rho r}, \quad w = \log\left(\frac{1+u}{1-u}\right), \quad t = \frac{u\sqrt{(n-2)}}{\sqrt{1-u^2}} \tag{78}$$

avem

Scanned with OKEN Scanner

PROPOZIȚIA 28 (Samiuddin (1970)). *Statistica u are probabilitatea elementară*

$$\frac{(n-2)(1-u^2)^{(n/2)-2}}{(n-1)\sqrt{2}\,B\left(\frac{1}{2}, n-\frac{1}{2}\right)}(1+u\rho)^{1/2}F\left\{\frac{1}{2}, \frac{1}{2}, n-\frac{1}{2}, \frac{1+2\rho u+\rho^2}{2(1+u\rho)}\right\}du$$

(79)

Pentru demonstrație se folosește (46).
Dezvoltînd F obținem:

$$M(t) = \frac{\rho(n-2)^{1/2}}{2(n-3)}\{1+0(1/n)\},$$

$$M(t^2) = \frac{n-2}{n-4}\{1+0(1/n)\},$$

(80)

$$M(t^3) = \frac{3\rho(n-2)^{3/2}}{2(n-3)^2}\{1+0(1/n)\},$$

$$M(t^4) = \frac{3(n-2)}{(n-4)(n-6)}\{1+0(1/n)\},$$

rezultate ce concordă cu cele obținute pentru o variabilă $t(n-2)$. Prin urmare, este rezonabil să presupunem că u are pe $[-1,1]$ densitatea

(81)
$$\frac{1}{B\left(\frac{1}{2}, \frac{1}{2}n-1\right)}(1-u^2)^{(n/2)-1}.$$

Samiuddin (1970) calculează funcția de repartiție pentru diferite valori ale lui r folosind aproximațiile

(i) u avînd densitatea (81)

(ii) $\frac{1}{2}\left(w - \frac{\rho}{n-1}\right)(n-3)^{1/2}$ ca $\sim N(0,1)$.

Avem:

$$(1-u^2)^{(n/2)-2} = (1-u)^{(n/2)-1}(1+u)^{(n/2)-2}$$

și

(82)
$$P(u \leqslant U) = \frac{1}{2}I_{U+1}\left(\frac{1}{2}n-1, \frac{1}{2}n-1\right),$$

ceea ce arată că (81) este ușor de interpolat.

Samiuddin dă probabilitatea P_1 folosind aproximația i) P_2 folosind aproximația ii) și probabilitatea exactă obținută din tabela lui David (1954)

Scanned with OKEN Scanner

Tabelul 2

r_1	r_2	P_1	P_2	Exactă
		(a) $n = 8$, $\rho = 0{,}9$		
−0,0500	0,9950	0,99825	0,99924	0,99845
−0,2893	0,9900	0,98835	0,99066	0,98925
−0,4635	0,9850	0,96936	0,96826	0,96969
−0,5696	0,9800	0,93537	0,93227	0,93876
		(b) $n = 8$, $\rho = 0{,}5$		
−0,7955	0,9759	0,99907	0,99974	0,99908
−0,6250	0,9500	0,99347	0,99575	0,99362
−0,3571	0,9000	0,99347	0,99575	0,99362
−0,1563	0,8500	0,89074	0,88473	0,89215
		(c) $n = 8$, $\rho = 0{,}3$		
−0,8375	0,95	0,99827	0,99933	0,99827
−0,5629	0,85	0,96353	0,96539	0,96374
−0,4459	0,80	0,92383	0,92168	0,92423
−0,2433	0,70	0,79955	0,78727	8,80028

Amîndouă aproximaţiile sînt rezonabile pentru selecţii de volum 8. Aproximaţia i) dă rezultate bune chiar pentru valori mari ale lui ρ şi arată că pentru valori mici ale lui ρ, concordanţa dintre P_1 şi valorile exacte devine strînsă.

Considerăm o selecţie de volum $N = n + 1$ asupra unui vector normal bidimensional cu coeficientul de corelaţie ρ.

O aproximaţie a repartiţiei statisticii r este dată în

LEMA 1. (Sadanori Konishi (1978)). *Fie $f(r)$ o funcţie analitică de coeficientul de corelaţie de selecţie r în $r = \rho$. Presupunem că $f'(\rho) \neq 0$ şi punem $m = n - 2\Delta$, unde Δ este un termen de corecţie ce depinde de ρ. Atunci dezvoltarea asimptotică pentru repartiţia lui $f(r)$ este:*

$$P\left\{m^{1/2}\frac{f(r)-f(\rho)}{(1-\rho^2)f'(\rho)}<x\right\}=\Phi(x)-m^{-1/2}\left\{\frac{1}{2}\rho-\left(\rho-\frac{1}{2}af''\right)x^2\right\}\varphi(x)+$$

$$+m^{-1}\left\{\left(\Delta-\frac{3}{4}+\frac{1}{8}\rho^2\right)x+\left(\frac{1}{4}+\frac{3}{2}\rho^2-\frac{5}{4}\rho af''+\frac{1}{2}a^2f''^2-\right.\right.$$

(83)

$$\left.\left.-\frac{1}{6}a^2f'f'''\right)x^3-\frac{1}{2}\left(\rho-\frac{1}{2}af''\right)^2x^5\right\}\varphi(x)+\mathbf{0}(m^{-3/2}),$$

unde $a = (1-\rho^2)f'^{-1}$ şi f', f'' şi f''' sînt derivatele funcţiei $f(r)$ cu $r = \rho$.

Observaţie. Pentru a obţine $P(r \leqslant r_0)(|r_0| < 1)$ din (83) luăm variabila x din (83) ca fiind

$$x = m^{1/2}\{f(r_0) - f(\rho)\}/\{(1 - \rho^2)f'(\rho)\}$$

şi $f(r)$ restrînsă la funcţii strict monotone în $|r| < 1$.

Scanned with OKEN Scanner

Prin urmare valoarea lui x în $r_0 = \rho$ este zero și devine mare cînd r_0 se depărtează de ρ, astfel că valorile lui x sînt mari în cozile repartiției. Prin urmare dorim să găsim o funcție care să reducă coeficientul lui x^2 la zero; adică

$$\rho - \frac{1}{2}(1 - \rho^2)f''(\rho)f'(\rho)^{-1} = 0.$$

Rezolvînd această ecuație diferențială și alegînd constanta convenabil obținem transformare z a lui Fisher.

$$f(r) = \frac{1}{2}\log\frac{1+r}{1-r}.$$

Aplicînd lema 1 la $f(r) = z(r)$ și luînd $\Delta = \frac{3}{4} - \frac{1}{8}\rho^2$, obținem aproximația S în forma:

$$P\left\{m^{1/2}\left(\frac{1}{2}\log\frac{1+r}{1-r} - \frac{1}{2}\log\frac{1+\rho}{1-\rho}\right) < x\right\} = \Phi(x) -$$

(83′)

$$-\frac{1}{2}\left(\rho m^{-1/2} + \frac{x^3}{6m}\right)\varphi(x) + \mathbf{0}(m^{-3/2}),$$

unde $m = n - \frac{3}{2} + \frac{1}{4}\rho^2$.

Cu alegerea $\Delta = \frac{3}{4} - \frac{1}{8}\rho^2$, termenul de ordinul $1/m$ are o formă simplă și aproximația dă o acuratețe înaltă pe întregul domeniu de definiție al statisticii r pentru valori mici ale lui N și diferite valori ale lui ρ. Nu a fost arătat că o astfel de alegere a lui Δ este optimă.

În cazul $f(r) = r$, (83) dă dezvoltarea asimptotică a funcției de repartiție a statisticii r, care conține totuși variabila x în termeni de ordinul $m^{-1/2}$. Aproximația normală pentru $z(r)$ cu media $z(\rho)$ și dispersia $1/(N-3)$ nu are acuratețe cu toate că această aproximație este superioară în coada repartiției cu repartiția normală limită a lui r.

Cu notațiile: R(aproximația lui Ruben (1966), K(aproximația t datorită lui Kraemer (1973)), F_1 (aproximația ce privește $z(r)$ ca o variabilă normală cu media $z(\rho)$ și dispersia $(N-3)^{-1}$) și F_2 (aproximația ce privește $z(r)$ ca o variabilă normală cu media $z(\rho) + \frac{1}{2}\rho/n + \frac{1}{8}\rho(5 + \rho^2)/n^2$ și dispersia $\left\{1 + \frac{1}{2}(4 - \rho^2)/n + \frac{1}{6}(22 - 6\rho^2 - 3\rho^4)n^2\right\}/n$, unde $n = N - 1$) avem (Sadanori Konishi (1978)) pentru $P(r \leqslant r_0)$

Tabelul 3

Max | Valoarea exactă − Valoarea aproximativă | $\times 10^5$, maximul fiind obținut cu ajutorul tabelelor lui David

ρ	N	F_1	F_2	R	K	S
0,1	11	1100	467	167	10	55
	25	549	198	29	8	10
	50	338	96	7	19	3
0,3	11	2278	481	269	75	82
	25	1321	199	63	36	17
	50	879	96	24	20	7
0,5	11	3551	498	400	219	98
	25	—	—	—	—	—
	50	—	—	—	—	—
0,7	11	4884	501	524	394	88
	25	2967	199	167	173	23
	50	—	—	—	—	—
0,9	11	6244	502	678	654	102
	25	3793	208	236	285	23
	50	2597	99	108	182	9

Valorile lipsă corespund la erorile posibile în tabelele lui David (1938)

Tabelul 3 arată că aproximația S prezintă o acuratețe mai bună decît celelalte aproximații pentru $\rho \geqslant 0{,}3$ și dă o acuratețe înaltă pe întregul domeniu al lui r, chiar pentru valori mici ale lui N. Aproximația K este exactă pentru $\rho = 0$.

§ 5. Covarianța de selecție

DEFINIȚIA 4. Statistica

$$n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \tag{84}$$

poartă numele de covarianța de selecție.

PROPOZIȚIA 29. *Dacă* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *este o selecție asupra vectorului* (X, Y) *avînd densitatea* (1), *atunci* (84) *este estimatorul de V.M. pentru* $\rho\sigma_1\sigma_2$.

DEFINIȚIA 5. *Statistica*

$$C = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \tag{85}$$

poartă numele de suma produselor abaterilor de la mediile de selecție.

Scanned with OKEN Scanner

Repartiţia statisticii C a fost determinată de Pearson, Jeffery, Elderton, (1929), Wishart, Bartlett (1932), Hirschfeld (1937) şi Mahalanobis, Bose, Roy (1937).

Din propoziţia 3 obţinem

PROPOZIŢIA 30. *Repartiţia statisticii C dat $x_1, x_2, \ldots, x_n$ este normală de valoare medie $(\rho\sigma_2/\sigma_1)S$ şi dispersie $\sigma_2^2\sqrt{(1-\rho^2)}S$ unde*

$$S = \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Ţinînd seama că $S \sim \sigma_1^2\chi^2(n-1)$ (corolarul III.3) are loc

PROPOZIŢIA 31. *Statistica C are densitatea de repartiţie*

$$p_C(c) = \frac{1}{\sigma_2\sqrt{2\pi(1-\rho^2)}} \int_0^\infty \frac{S^{(n-4)/2} \exp\left(-\frac{1}{2} S/\sigma_1^2\right)}{(2\sigma_1^2)^{(n-1)/2}\, \Gamma\left(\frac{1}{2}(n-1)\right)} \times$$

$$\times \exp\left[-\frac{(c - \rho\sigma_2 S/\sigma_1)^2}{2\sigma_2^2(1-\rho^2)S}\right] dS.$$

Observaţie. Dacă $\sigma_1 = \sigma_2 = 1$, atunci

$$p_C(c) = \frac{e^{c\rho/(1-\rho^2)}}{2^{n/2}\, \Gamma\left(\frac{1}{2}(n-1)\right)\sqrt{\pi(1-\rho^2)}} \times$$

$$\times \int_0^\infty S^{(n-4)/2} \exp\left[-\frac{S + c^2 S^{-1}}{2(1-\rho^2)}\right] dS =$$

$$= \frac{|c|^{(n/2)-1}\, e^{c\rho/(1-\rho^2)}}{2^{(n/2)-1}\, \Gamma\left(\frac{1}{2}(n-1)\right)\sqrt{\pi(1-\rho^2)}} K_{\frac{1}{2}(n-2)}\left(\frac{|c|}{1-\rho^2}\right),$$

unde $K_\nu(z)$ este funcţia Bessel de tipul doi, de ordinul ν.

Press (1967) observă că dacă n este impar, atunci este posibil să se exprime densitatea ca o serie finită de funcţii elementare şi ţine seama de :

$$K_{(n/2)-1}(z) = \sqrt{\frac{\pi}{2z}}\, e^{-z} \sum_{j=0}^{\frac{1}{2}(n-3)} \frac{\left(\frac{1}{2}(n-3)+j\right)!}{j!\left(\frac{1}{2}(n-3)-j\right)!} \cdot \frac{1}{(2z)^j}.$$

Scanned with OKEN Scanner

Funcţia caracteristică a lui C este

$$\{1 - 2it\rho\sigma_1\sigma_2 + t^2(1-\rho^2)\sigma_1^2\sigma_2^2\}^{-(n-1)/2} =$$

$$= \left[\left\{1 - 2it\frac{\sigma_1\sigma_2(1+\rho)}{2}\right\}\left\{1 + 2it\frac{\sigma_1\sigma_2(1-\rho)}{2}\right\}\right]^{-(n-1)/2}. \tag{88}$$

Deci repartiţia lui C este aceeaşi ca a statisticii

$$\frac{1}{2}\sigma_1\sigma_2\{(1+\rho)Z_1 - (1-\rho)Z_2\}$$

unde Z_1 şi Z_2 sînt variabile aleatoare independente, fiecare repartizată $\chi^2(n-1)$

Ţinînd seama de reprezentarea

$$C = \frac{1}{2}\sigma_1\sigma_2\{(1+\rho)Z_1 - (1-\rho)Z_2\} \tag{89}$$

obţinem

PROPOZIŢIA 32. *Avem*

$$\mathfrak{M}_r(C) = \left(\frac{1}{2}\sigma_1\sigma_2\right)^r\{(1+\rho)^r + (-1)^r(1-\rho)^r\}\mathfrak{M}_r(\chi^2(n-1)) \tag{90}$$

şi în particular

$$M(C) = (n-1)\rho\sigma_1\sigma_2. \tag{91}$$

$$D^2(C) = (n-1)(1+\rho^2)\sigma_1^2\sigma_2^2 \tag{92}$$

$$\alpha_3(C) = \sqrt{\beta_1(C)} = 2\rho(3+\rho^2)(1+\rho^2)^{-3/2}(n-1)^{-1/2} \tag{93}$$

$$\alpha_4(C) = \beta_2(C) = 3 + 6(1 + 6\rho^2 + \rho^4)(1+\rho^2)^{-2}(n-1)^{-1}. \tag{94}$$

Observaţie. Reprezentarea (89) arată că pentru $n \to \infty$, repartiţia statisticii C tinde către repartiţia normală.

§ 6. Corelaţie serială circulară

În formula (39) a coeficientului de corelaţie, în mod obişnuit X_t şi Y_t reprezintă observaţii asupra aceluiaşi individ privind 2 caracteristici deosebite. Cu toate că această situaţie apare în foarte multe aplicaţii, nu este necesară formal şi unele tehnici pot fi bazate pe modificări ale acestei situaţii. În particular se poate lua $Y_t = X_{t+k}$.

DEFINIŢIA 6. Coeficienţii de corelaţie (39) cu $Y_t = X_{t+k}$ poartă numele de corelaţii seriale.

Scanned with OKEN Scanner

DEFINIȚIA 7. Valoarea absolută a diferenței dintre t și indicele lui X corespunzător lui Y_t poartă numele de lag-ul corelației.

Observație i) Sînt obținute rezultate similare luînd $Y_t = X_{t+k}$ sau $Y_t = X_{t-k}$

ii) Raportul patratelor diferențelor succesive (von Neumann (1941))

$$d = \frac{\sum_{t=1}^{n-1} (X_t - X_{t+1})^2}{\sum_{t=1}^{n} (X_t - \bar{X})^2} \tag{95}$$

este strîns legat de corelația serială (R_1) de lag1. În fapt

$$d = \frac{\sum_{t=1}^{n-1} (X_t - \bar{X})^2 - 2\sum_{t=1}^{n-1} (X_t - \bar{X})(X_{t+1} - \bar{X}) + \sum_{t=1}^{n-1} (X_{t+1} - \bar{X})^2}{\sum_{t=1}^{n} (X_t - \bar{X})^2},$$

$$= 2(1 - R_1) - \frac{(X_1 - \bar{X})^2 + (X_n - \bar{X})^2}{\sum_{t=1}^{n} (X_t - \bar{X})^2}. \tag{96}$$

Un mod de a recunoaște efectele de sfîrșit este de a privi datele ca constînd din $(n - k)$ perechi.

$$(X_1, X_{1+k}), (X_2, X_{2+k}), \ldots, (X_{n-k}, X_n).$$

DEFINIȚIA 8. Corelația serială bazată pe perechile

$$(X_t, X_{t+k}) \; t = 1, \ldots, n$$

poartă numele de corelație serială circulară.

Punem $X_{n+j} = X_j$ pentru $j = 1, 2, \ldots, k$. În acest caz, numitorul lui R_k este $\sum_{t=1}^{n} (X_t - \bar{X})^2$. Dacă k este mic comparativ cu n, repartițiile corespunzătoare corelațiilor circulare pot fi aplicate în corelațiile necirculare cu un risc mic în obținerea erorilor serioase.

În cazul raportului pătratelor diferențelor succesive, definiția circularității conduce la

$$\tilde{d} = \sum_{t=1}^{n} (X_t - X_{t+1})^2 / \sum_{t=1}^{n} (X_t - \bar{X})^2 \tag{95'}$$

în timp ce corelația serială circulară de lag1 este

$$\tilde{R}_1 = \sum_{t=1}^{n} (X_t - \bar{X})(X_{t+1} - \bar{X}) / \sum_{t=1}^{n} (X_t - \bar{X})^2 \tag{39'}$$

Scanned with OKEN Scanner

(cu $X_{n+1} = X_1$). Relația dintre $\tilde{d}$ și $\tilde{R}_1$ este

$$\tilde{d} = 2(1 - \tilde{R}_1).$$

Considerăm repartiția corelației seriale circulare de lag1 cînd X-șii sînt legați prin ecuații de forma

$$X_t = \rho X_{t-1} + Z_t (|\rho| < 1;\ t = 1,2, \ldots, n) \tag{97}$$

unde $\{Z_t\}$ sînt variabilele aleatoare independente, fiecare avînd o repartiție $N(0,\ \sigma^2)$ și în plus Z_t este independentă de toți X_j pentru $j < t$. Se presupune că seria pornește cu o variabilă aleatoare X_0 care este repartizată $N(0,\ \sigma^2(1 - \rho^2)^{-1})$. Efectul acestei presupuneri este de a asigura ca fiecare X_t să aibă aceeași dispersie (σ^2) cît și aceeași valoare medie (zero).

Atît R_1 cît și $\tilde{R}_1$ sînt cazuri particulare de rapoarte de forme pătratice în variabile normale centrate.
Metodele folosite de von Neumann (1941) în obținerea repartiției lui d (pentru $\rho = 0$) pot fi folosite de asemenea în obținerea repartițiilor corelațiilor seriale (pentru $\rho = 0$ sau $\rho \neq 0$). Rezultatele au proprietatea caracteristică că densitățile iau forme diferite pentru intervale diferite în domeniul de variație ale variabilei.

Teorema 1 (Anderson (1942) (se poate vedea de asemenea Koopmans (1942)). *Dacă* $\rho = 0$, *atunci*

$$P(\tilde{R}_1 > r) = \sum_{j=1}^{n} \alpha_j^{-1} (\lambda_j - r)^{(n-3)/2} \text{ (pentru } \lambda_{m+1} \leqslant r \leqslant \lambda_m,\ m = 1,2, \ldots, n-1) \tag{98}$$

unde $\lambda_j = \cos(2\pi j/n)$

$$\alpha_j = \begin{cases} \prod_{i \neq j,\ i=1}^{(n-1)/2} (\lambda_j - \lambda_i) & , \ n \text{ impar} \\ \left[\prod_{i \neq j,\ i=1}^{(n-2)/2} (\lambda_j - \lambda_i) \right] \sqrt{1 + \lambda_j}, & n \text{ par} \end{cases}$$

Anderson a obținut de asemenea repartiția statisticii R (corelația serială circulară cu lag-ul l). Anderson consideră de asemenea repartiția statisticii

$$\tilde{R}_1' = \frac{\sum_{j=1}^{n} X_j X_{j+1}}{\sum_{j=1}^{n} X_j^2} \tag{99}$$

(unde se folosește faptul că $M(X_j) = 0$).
Are loc

Teorema 2. *Dacă* α_j *este multiplu de* $\sqrt{1 - \lambda_j}$, n *este impar și* $\rho = 0$, *repartiția statisticii* $\tilde{R}_1'$ *este aceeași cu a statisticii* $\tilde{R}_1$ *cu* n *crescut cu* 1.

Scanned with OKEN Scanner

Madow (1945) extinde aceste rezultate la cazul $\rho \neq 0$. În acest scop foloseşte faptul că dacă $(T_1, T_2, \ldots, T_s) \equiv \mathbf{T}$ este o mulţime suficientă de statistici pentru parametrul θ, atunci

$$\frac{p_{\mathbf{T}}(t|\rho)}{p_{\mathbf{T}}(t|0)} = \frac{p_{x_1,\ldots,x_n}(x_1, \ldots, x_n|\rho)}{p_{x_1,\ldots,x_n}(x_1, \ldots, x_n|0)}$$

astfel că repartiţia lui $\mathbf{T}$ pentru ρ general poate fi obţinută din repartiţia pentru $\rho = 0$.

TEOREMA 3 (Madow (1945)). $\tilde{R}_1$ are densitatea de repartiţie

$$(100)\quad p_{\tilde{R}_1}(r) = \frac{1}{2}(n-3)\frac{\prod_{j=1}^{n}(1+\rho^2-\lambda_j\rho)^{1/2}}{(1+\rho^2-2\rho r)^{(n/2)-1}}\sum_{j=1}^{m}\alpha_j^{-1}(\lambda_j - r)^{(n-5)/2}$$

pentru $\lambda_{m+1} \leqslant r \leqslant \lambda_m$, $m = 1, \ldots, (n-1)$.

TEOREMA 4. (Daniels (1956)). *Densitatea de repartiţie a statisticii $\tilde{R}_1'$ admite dezvoltarea:*

$$(101)\quad p_{\tilde{R}_1'}(r) = \frac{\Gamma\left(\frac{1}{2}n+1\right)(1-\rho^n)}{\sqrt{\pi}\,(1+\rho^2-2\rho r)^{n/2}} \times$$
$$\times\left\{\frac{(1-r^2)^{(n-1)/2}}{\Gamma\left(\frac{1}{2}n+\frac{1}{2}\right)} - \frac{3}{2^r\Gamma\left(\frac{3}{2}n+\frac{1}{2}\right)}\frac{d^n}{dr^n}(1-r^2)^{(3n-1)/2} + \ldots\right\}$$

termenul j fiind:

$$(-1)^{j-1}\frac{2j-1}{2^{jn}\Gamma\left(\left(j-\frac{1}{2}\right)n+\frac{1}{2}\right)}\cdot\frac{d^{jn}}{dr^{jn}}(1-r^2)^{[(2j-1)n-1]/2}.$$

Repartiţia aleatorizată ($\rho = 0$) a statisticii $\tilde{R}_1$ (sau $\tilde{R}_1'$) este repartiţia obţinută considerînd toate ordonările posibile ale celor n valori observate $X_1, \ldots, X_n$. Deoarece numitorul nu se schimbă prin reordonare, această repartiţie este repartiţia a n valori posibile ale numărătorului. Această repartiţie a fost studiată de Wald, Wolfowitz (1943).

§ 7. Corelaţie serială necirculară

Repartiţiile corelaţiilor seriale necirculare sînt în general mult mai complicate decît acelea corespunzătoare corelaţiilor circulare. Este posibil să obţinem formule pentru momente.

Scanned with OKEN Scanner

Mai întîi considerăm corelația serială circulară de lag 1 ca

$$\hat{R}_1 = \left(\sum_{j=1}^{n} X_{j-1}X_j\right) \Big/ \left(\sum_{j=1}^{n-1} X_j^2\right) \tag{102}$$

cu $X_0 = 0$ și considerăm repartiția sa în modelul (97). Densitatea de repartiție comună a variabilelor $X_1, \ldots, X_n$ este

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = (2\pi\sigma)^{-\frac{1}{2}n} \exp\left[-\sum_{j=1}^{n} (x_j - \rho x_{j-1})^2\right].$$

Repartiția statisticii $\hat{R}_1$ nu depinde de σ, ceeace ne permite să luăm $\sigma = 1$.

Funcția generatoare de momente comună a numărătorului și numitorului statisticii $\hat{R}_1$ este

$$M\left[\theta_1 \sum_{j=1}^{n} X_{j-1}X_j + \theta_2 \sum_{j=1}^{n-1} X_j^2\right] = [D_n(\theta_1, \theta_2)]^{-1/2}$$

unde, pentru $n \geqslant 2$

$$D_n(\theta_1, \theta_2) = (1 + \rho^2 - 2\theta_2)D_{n-1}(\theta_1, \theta_2) - (\rho + \theta_1)^2 D_{n-2}(\theta_1, \theta_2) \tag{103}$$

cu $D_0(\theta_1, \theta_2) = D_1(\theta_1, \theta_2) = 1$.

White (1957) evaluează momentele de primul și al doilea ordin pentru $\hat{R}_1$ din

$$\mu'_m(\hat{R}_1) = \int_{-\infty}^{0}\int_{-\infty}^{t_m} \cdots \int_{-\infty}^{t_2} \frac{\partial^m \{D_n(\theta_1, t_1)\}^{-1/2}}{\partial \theta_1^m}\Bigg|_{\theta_1=0} dt_1 dt_2 \ldots dt_m . \tag{104}$$

TEOREMA 5. (Shenton, Johnson (1965)). *Momentul lui $\hat{R}_1$ față de ρ sînt date de*

$$M[(\hat{R}_1 - \rho)^m] = \sum_{j=0}^{[m/2]} a_j^{(m)} \int_0^\infty \left(\frac{d}{d\rho}\right)^{m-2j} H_{m-j}(t)\, dt \tag{105}$$

unde

$$H_r(t) = \frac{t^r}{r!}\{D_n(0, t)\}^{-1/2},$$

$$\sum_{j=0}^{[m/2]} (-1)^j a_j^{(m)} y^{m-2j} = e^{-\frac{1}{2}D^2} y^m \text{ *)} \quad \text{și} \quad a_0^{(m)} = 1.$$

TEOREMA 6 (Sheton, Johnson (1965)). *În forma unor serii asimptotice avem:*

$$M(\hat{R}_1) = \rho - \frac{2(n-2)\rho}{(n+1)(n-1)} - \frac{12\rho^3}{(n+5)(n+3)(n+1)} + \frac{18(n+8)\rho^5}{(n+9)^{[4]}} + \tag{106}$$

*) D reprezintă operatorul diferențial.

$$+\frac{24(n+10)(n+12)\rho^7}{(n+13)^{[5]}}+\frac{30(n+12)(n+14)(n+16)\rho^9}{(n+17)^{[6]}}+\ldots$$

$$M[(\hat{R}_1-\rho)^2]=\frac{n^2-4n+7}{(n+1)^{[3]}}-\frac{(n^3-6n^2-25n+42)\rho^2}{(n+5)^{[4]}}+$$

(107)

$$+\frac{3(n+19)(n-3)\rho^4}{(n+9)^{[5]}}+\frac{3(n^3+32n^2+111n-928)\rho^6}{(n+5)^{[6]}}+$$

$$+\frac{3(n+12)(n^3+42n^2+215n-1950)\rho^8}{(n+17)^{[7]}}+\ldots$$

unde $n^{[s]}=n(n-2)(n-4)\ldots(n-2\cdot\overline{s-1})$.

Pentru valori mari ale lui n au loc următoarele dezvoltări în puteri de n^{-1}

$$M(\hat{R}_1)=\rho-\frac{2\rho}{n}+\frac{4\rho}{n^2}-\frac{2\rho(1-8\rho^2+4\rho^4)}{n^3(1-\rho^2)^2}+$$

(108)

$$+\frac{4\rho(1-30\rho^2+12\rho^4-4\rho^6)}{n^4(1-\rho^2)^3}-\ldots$$

$$M[(\hat{R}_1-\rho)^2]=\frac{1-\rho^2}{n}-\frac{1-14\rho^2}{n^2}+\frac{5-78\rho^2+76\rho^4}{n^3(1-\rho^2)}+$$

(109)

$$+\frac{11+316\rho^2-692\rho^4+341\rho^6}{n^4(1-\rho^2)^2}+\ldots$$

$$D^2(\hat{R}_1)=\frac{1-\rho^2}{n}-\frac{1-10\rho^2}{n^2}+\frac{5-62\rho^2+60\rho^4}{n^3(1-\rho^2)}+$$

(110)

$$+\frac{11+292\rho^2-596\rho^4+296\rho^6}{n^4(1-\rho^2)^2}+\ldots$$

Shenton, Johnson (1965) au obținut de asemenea formule pentru β_1 și β_2, pe care nu le publică însă dau rezultate bazate pe acestea (incluse în tabela 4).

S-au folosit notațiile: (a) $M(\hat{R}_1-\rho)$, (b) $D(\hat{R}_1)$, (c) $\beta_1(\hat{R}_1)$, (d) $\beta_2(\hat{R}_1)$; $M(\hat{R}_1-\rho)$ și $\beta_1(\hat{R}_1)$ sînt zero pentru $\rho=0$; și valorile marcate cu semnul + sînt afectate de erori de rotunjire.

Scanned with OKEN Scanner

Pan (1964) obține repartiția exactă a statisticii $\hat{R}_1$ în cazul în care corelația dintre X_i și X_j este ρ pentru $|i-j|=1$ și zero în rest. În acest caz $\sum_{j=1}^{n}(X_j-\bar{X})^2$ este repartizată ca un multiplu de χ^2 și este posibil să se urmeze metoda lui von Neumann (1941).

Tabelul 4

Valorile momentelor repartiției statisticii $\hat{R}_1$ (Shenton, Johnson 1965)

ρ	n	6	8	10	15	50	100	500
0,0	(b)	0,4254	0,3519	0,3109	0,2530	0,1401	0,0995	0,0447
	(d)	4,7331	3,0022	2,7859	2,7440	2,8919	2,9430	2,9881
0,2	(a)	0,0456	0,0380	0,0323	0,0232	0,0077	0,0039	0,0008
	(b)	0,4244	0,3500	0,3086	0,2502	0,1378	0,0977	0,0438
	(c)	0,0267	0,0418	0,0476	0,0477	0,0241	0,0134	0,0029
	(d)	4,8014	3,0767	2,8625	2,8171	2,9301	2,9648	2,9930
0,4	(a)	0,0902	0,0755	0,0642	0,0463	0,0154	0,0078	0,0016
	(b)	0,4217	0,3445	0,3017	0,2419	0,1304	0,0920	0,0410
	(c)	0,1067	0,1719	0,1999	0,2067	0,1085	0,0610	0,0134
	(d)	5,0031	3,3104	3,1096	3,0613	3,0647	3,0422	3,0103
0,6	(a)	0,1321	0,1114	0,0952	0,0689	0,0230	0,0118	0,0024
	(b)	0,4182	0,3358	0,2901	0,2275	0,1172	0,0815	0,0359
	(c)	0,2392	0,4021	0,4872	0,5372	0,3099	0,1774	0,0394
	(d)	5,3160	3,7301	3,5833	3,5721	3,3839	3,2310	3,0535
0,8	(a)	0,1674	0,1426	0,1228	0,0900	0,0306	0,0157	0,0032
	(b)	0,4155	0,3256	0,2751	0,2067	0,0958	0,0641	0,0272
	(c)	0,4319	0,7466	0,9563	1,1876	0,9171	0,5318	0,1235
	(d)	5,6832	4,3496	4,3732	4,5969	3,8334	3,8032	3,1926
1,0	(a)	0,1893	0,1599	0,1369	0,0998	0,034+	0,019+	0,004+
	(b)	0,4146	0,3167	0,2603	0,1826	0,06+	0,03+	0,006+
	(c)	0,7456	1,2859	1,7182	2,4705	3,89+	5,1+	6,5+
	(d)	6,0955	5,1633	5,5432	6,5897	9,3+	11,3+	13,2+

TEOREMA 7. *Statistica $\hat{R}_1$ este repartizată ca*

$$\left(\sum_{j=1}^{n-1}\lambda_j U_j^2\right)\Big/\left(\sum_{j=1}^{n-1}U_j^2\right) \tag{111}$$

unde $U_1, U_2, \ldots, U_{n-1}$ sînt variabile aleatoare independente repartizate $N(0,1)$ și
$\lambda_1 > \lambda_2 > \ldots > \lambda_{n-1}$
și zero sînt rădăcinile caracteristice ale matricei **MAM** *cu*

$$\mathbf{M} = \begin{pmatrix} 1-n^{-1} & -n^{-1} & \ldots & -n^{-1} \\ n^{-1} & 1-n^{-1} & \ldots & -n^{-1} \\ \vdots & \vdots & & \vdots \\ n^{-1} & n^{-1} & \ldots & 1-n^{-1} \end{pmatrix} \tag{112}$$

Scanned with OKEN Scanner

(113) $$\mathbf{A} = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \dots & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \dots & 0 & 0 \end{pmatrix}.$$

Valorile lui λ, pentru ordin impar, sînt

(114) $$\lambda_{2k-1}| = \cos\frac{2k\pi}{n+1} \quad \left(k = 1, 2, \dots, \left[\frac{1}{2}n\right]\right)$$

în timp ce $\lambda_2, \lambda_4, \dots, \lambda_{2[(n-1)/2]}$ sînt rădăcinile ecuației

(115)] $$(1-\lambda)^{-2}\left\{\left(-\frac{1}{2}\right)^n - \frac{1}{2}D_{n-1}(\lambda) - (n+1-n\lambda)D_n(\lambda)\right\}$$

cu

(116) $$D_n(\lambda) = \left(-\frac{1}{2}\lambda\right)^n \sum_{j=1}^{\left[\frac{1}{2}n+1\right]} C_{n+1}^{2j-1}(-1)^{j-1}(\lambda^2-1)^{j-1}.$$

TEOREMA 8. *Funcția de repartiție a statisticii* $\hat{R}_1$ *are forme diferite în fiecare din intervalele* $\lambda_{j+1} \leqslant \hat{R}_1 \leqslant \lambda_j$, *fiind* $(\lambda_{j+1} \leqslant r \leqslant \lambda_j)$

(117)
$$F_{\hat{R}_1}(r) = 1 + \pi^{-1}\sum_{i=1}^{[j/2]}(-1)^i \int_{-1}^{1} \frac{(y_{2i}(t)-r)^{(n-3)/2}}{\prod_{\substack{k=1\\k\neq j}}^{n-1}|y_j(t)-\lambda_k|^{1/2}} \cdot \frac{dt}{\sqrt{1-t^2}} +$$
$$+ \frac{1-(-1)^j}{\pi}(-1)^{(j+1)/2}\int_{-1}^{1} \frac{(y_j(t)-r)^{(n/2)-1}}{\prod_{\substack{k=1\\k\neq j}}^{n-1}|y_j(t)-\lambda_k|^{1/2}} \cdot \frac{dt}{\sqrt{1-t^2}}.$$

Pentru valori ale lui N suficient de mari integralele din (117) pot fi evaluate cu ajutorul formulei

$$\int_{-1}^{1} f(y)\frac{dy}{\sqrt{1-y^2}} \simeq \pi N^{-1}\sum_{j=1}^{N} f(y_j^{(N)})$$

cu

$$y_j^{(N)} = \cos\frac{(2j-1)\pi}{2N}.$$

Scanned with OKEN Scanner

Observație. Variabilele $U_j^2\left[\sum_{j=1}^{n-1} U_j^2\right]^{-1}$ sînt variabile beta corelate cu o repartiție comună Dirichlet și astfel $\hat{R}_1$ este o funcție liniară de astfel de variabile.

Pan (1964) a considerat de asemenea o modificare a coeficientului de corelație serial necircular obținut prin împărțirea șirului de $2n$ valori $X_1, \ldots, X_{2n}$ în 2 mulțimi, primele n și respectiv ultimele n valori. Coeficientul este definit ca :

$$R_{1,1} = \frac{\sum_{j=1}^{n-1}(X_j - \bar{X}_1)(X_{j+1} - \bar{X}_1) + \sum_{j=n+1}^{2n-1}(X_j - \bar{X}_2)(X_{j+1} - \bar{X}_2)}{\sum_{j=1}^{n}(X_j - \bar{X}_1)^2 + \sum_{j=n+1}^{2n}(X_j - \bar{X}_2)^2}, \tag{118}$$

unde

$$\bar{X}_1 = n^{-1}\sum_{j=1}^{n} X_j; \quad \bar{X}_2 = n^{-1}\sum_{j=n+1}^{2n} X_j.$$

TEOREMA 9. *Statistica $R_{1,1}$ este repartizată ca*

$$\left(\sum_{j=1}^{n-1} \lambda_j V_j\right) \Big/ \left(\sum_{j=1}^{n-1} V_j\right) \tag{119}$$

unde V_j sînt variabile aleatoare independente, fiecare repartizată ca $\chi^2(2)$. λ_j au aceleași valori ca pentru $\hat{R}_1$.

TEOREMA 10. *Funcția de repartiție a statisticii $R_{1,1}$ este*

$$P(R_{1,1} \leqslant r) = 1 - \sum_{i=1}^{j}\left[\prod_{\substack{k=1\\k\neq j}}^{n-1}(\lambda_j - \lambda_k)\right]^{-1}(\lambda_j - r)^{n-2} \tag{120}$$

pentru $\lambda_{j+1} \leqslant r \leqslant \lambda_j$.

Cercetări similare privind repartiția statisticii $R_{1,1}$ au fost făcute de Watson, Durbin (1951) care au dat de asemenea tabele cu punctele de semnificație de 5% pentru un test exact necircular privind existența unei corelații seriale într-o serie de n observații.

Pan (1965) dă repartiția asimptotică a statisticii $\hat{R}_l$ pentru l general.

Repartiția aleatoare ($\rho = 0$) a statisticii $\hat{R}_1$ a fost studiată de David, Fix (1966).

§8. Repartiția Leipnik

Repartițiile descrise în §5 și §6 au forme complicate. Dixon (1944), Rubin (1945) și apoi Leipnik (1947) propun o metodă de *netezire* a funcției caracteristice (eliminînd discontinuitățile lui $p_{\tilde{R}_1'}(r)$) și obțin

$$p_{\tilde{R}_1'}(r) = \left[B\left(\frac{1}{2}, \frac{1}{2}(n+1)\right)\right]^{-1}(1-r^2)^{(n-1)/2}(1+\rho^2-2\rho r)^{-n/2}, \tag{121}$$

$$(-1 \leqslant r \leqslant 1).$$

Observații. i) Acesta este primul termen al dezvoltării (101), omițînd $-\rho^n$.

ii) Daniels (1956) cercetează eroarea acestei aproximații și obține o margine superioară pentru ea.

iii) Dacă $\rho = 0$, avem

$$p_{\tilde{R}_1'}(r) = \left[B\left(\frac{1}{2}, \frac{1}{2}(n+1)\right)\right]^{-1}(1-r^2)^{(n-1)/2}$$

care reprezintă densitatea corelației obișnuite (neseriale) pentru o populație normală bidimensională cu $\rho = 0$ cînd volumul selecției este $(n+3)$.

White (1957) exprimă $(1+\rho^2-2\rho r)^{-n/2}$ în termeni de polinoame Gegenbauer, $c_j^{(n/2)}(x)$ ca

$$\sum_{j=0}^{\infty} c_j^{(n/2)}(r)\,\rho^j$$

unde

$$c_j^{(n/2)}(x) = \sum_{m=0}^{[1/2j]} \frac{\left(-\frac{1}{2}n\right)^{(j-m)}(-2x)^{j-2m}}{m!\,(j-2m)!}.$$

Avem

$$x^{(r)} = x(x-1)\ldots(x-r+1)$$

și astfel

$$(-x)^{(r)} = (-x)(-x-1)\ldots(-x-r+1) =$$

$$= (-1)^r x(x+1)\ldots(x+r-1).$$

Polinoamele Gegenbauer sînt un caz special al polinoamelor Jacobi. Ele sînt ortogonale în raport cu funcția de pondere $(1-x^2)^{(n-1)/2}$ pe intervalul $-1 \leqslant x \leqslant 1$.

Propoziția 33. *Funcția de repartiție a statisticii* $\tilde{R}_1'$ *este*

$$P(\tilde{R}_1' \leqslant R_0) = I_{(R_0+1)/2}\left(\frac{n+1}{2}, \frac{n+1}{2}\right) -$$

(122)

$$-\frac{n}{B\left(\frac{1}{2}, \frac{n+1}{2}\right)}(1-R_0^2)^{(n+1)/2}\sum_{j=1}^{\infty}\frac{\rho^j}{j(n+j)}c_{j-1}^{(n+2)/2}(R_0).$$

Propoziția 34. *Funcția generatoare de momente a statisticii* $\tilde{R}_1'$ *este*

$$M(e^{it\tilde{R}_1'}) = (2t^{-1})^{n/2}\Gamma\left(\frac{1}{2}n+1\right)\sum_{j=0}^{\infty}(i\rho)^j\frac{r(n+j)}{j!\,\Gamma(n)}J_{\frac{1}{2}n+j}(t) \tag{123}$$

unde $J_\nu(t)$ *este funcția Bessel de ordinul* ν.

Scanned with OKEN Scanner

PROPOZIȚIA 35 (Jenkins (1956)). *Momentul de ordinul r față de zero al statisticii* $\tilde{R}_1'$ *este*

$$\mu_r' = \frac{r!}{2^r} \frac{\Gamma\left(\frac{n}{2}+1\right)}{\Gamma(n)} \sum_{j=0}^{[r/2]} \frac{\Gamma(n+r-2j)}{\Gamma\left(\frac{1}{2}n+r-j+1\right)} \cdot \frac{\rho^{r-2j}}{j!\,(r-2j)!}. \tag{124}$$

Din (124), găsim

$$\mu_1' = \frac{n}{n+2}\rho \tag{125}$$

$$\mu_2 = \frac{1}{n+2} - \frac{n(n-2)\rho^2}{(n+2)^2(n+4)} \tag{126}$$

$$\mu_3 = \frac{1}{(n+2)^2}\left[-\frac{6n\rho}{n+4} + \frac{2n(n-2)(3n-2)\rho^3}{(n+2)(n+4)(n+6)}\right] \tag{127}$$

$$\mu_4 = \frac{12}{(n+2)(n+4)}\left[1 - \frac{2n(n^2-8n-4)\rho^3}{(n+2)^2(n+6)} + \right. \tag{128}$$
$$\left. + \frac{n(n^4-16n^3+40n^2-32n+16)\rho^4}{(n+2)^3(n+6)(n+8)}\right].$$

Ne putem aștepta ca transformarea de egalizare a dispersiei arctgh $\tilde{R}_1'$ să producă o variabilă mai aproape de normalitate care este cazul pentru (39). O astfel de transformare a fost studiată de Quenouille (1948).

Totuși din (126) se vede că

$$D^2(\tilde{R}_1') \simeq \frac{1-\rho^2}{n} \tag{129}$$

în timp ce

$$D^2(R) \simeq \frac{(1-\rho^2)^2}{n}.$$

Ecuația (129) sugerează transformarea

$$\tilde{Z} = \arcsin \tilde{R}_1'. \tag{130}$$

PROPOZIȚIA 36 (Jenkins (1954)). (130) *este comparabil în eficiență pentru* $\tilde{R}_1'$, *cu* z' *pentru* R.

Scanned with OKEN Scanner

PROPOZIȚIA 37 (Jenkins (1954)). *Avem*

$$M(\tilde{Z}) = \arcsin\rho - \frac{3}{2}\rho(1-\rho^2)^{-1/2}n^{-1} +$$

(131)

$$+\frac{1}{8}\rho(17-2\rho^2)(1-\rho^2)^{-3/2}n^{-2}+\mathbf{0}(n^{-3})$$

(132) $$D^2(\tilde{Z}) = n^{-1} - \frac{1}{2}(2-5\rho^2)(1-\rho^2)^{-1}n^{-2} + \mathbf{0}(n^{-3})$$

(133) $$\mu_3(\tilde{Z}) = -3\rho(1-\rho^2)^{-1/2}n^{-2}+\mathbf{0}(n^{-3})$$

(134) $$\mu_4(\tilde{Z}) = 3n^{-2} - (8-29\rho^2)(1-\rho^2)^{-1}n^{-3} + \mathbf{0}(n^{-4})$$

(135) $$\sqrt{\beta_1} = -3\rho(1-\rho^2)^{-1/2}n^{-1/2}\left\{1+\frac{3}{4}(2-5\rho^2)(1-\rho^2)^{-1}n^{-1}\right\} + \mathbf{0}(n^{-5/2})$$

(136) $$\beta_2 = 3 + 2(7\rho^2-1)(1-\rho^2)^{-1}n^{-1} + \mathbf{0}(n^{-2}).$$

§9. Coeficientul de corelație multiplă

DEFINIȚIA 9. Coeficientul de corelație multiplă dintre o variabilă aleatoare X_0 (*variabila dependentă*) și variabilele $X_1, X_2, \ldots, X_k$ (variabilele independente) cu $k \geqslant 2$ este definit ca fiind corelația maximă dintre X_0 și orice funcție liniară de variabilele *independente*, adică

(137) $$P_{0.12\ldots k} = \max_{a_1, a_2, \ldots, a_k} \rho\left(X_0, \sum_{j=1}^{k} a_j X_j\right).$$

Observație. Unde nu este nici un risc de confuzie indicii $0.12\ldots k$ pot fi omiși.

Dacă matricea de covarianță a variabilelor $X_0, X_1, \ldots, X_k$ este

$$\mathbf{V} = \begin{pmatrix} D^2(X_0) & \mathbf{V}_0' \\ \mathbf{V}_0 & \mathbf{V}_{(1)} \end{pmatrix}$$

unde $\mathbf{V}_{(1)}$ este matricea de covarianță a variabilelor $X_1, \ldots, X_k$, atunci

(138) $$\rho\left(X_0, \sum_{j=1}^{k} a_j X_j\right) = (\mathbf{V}_0'\mathbf{a})/\{(\mathbf{a}'\mathbf{V}_{(1)}\mathbf{a})D^2(X_0)\}^{1/2}.$$

Deoarece printr-o alegere convenabilă a semnelor lui $a_1, \ldots, a_k$, se poate face ca $\rho(\cdot)$ să fie întotdeauna nenegativ și se poate alege „$\mathbf{a}$" să maximizeze

$$(\mathbf{a}'\mathbf{V}_0\mathbf{V}_0'\mathbf{a})/(\mathbf{a}'\mathbf{V}_{(1)}\mathbf{a}).$$

Scanned with OKEN Scanner

Valoarea maximizată a pătratului coeficientului de corelaţie este

$$\mathbf{V}_0'\mathbf{V}^{-1}\mathbf{V}_0/D(X_0),$$

astfel că coeficientul de corelaţie multiplă este

$$P = \sqrt{\mathbf{V}_0'\mathbf{V}^{-1}\mathbf{V}_0/D(X_0)}. \tag{139}$$

Reluăm definiţia 9 într-un caz particular.

Fie $\mathbf{x}$ un vector m-dimensional $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ portiţionat în vectorii: $\mathbf{x}^1$, k dimensional şi $\mathbf{x}^2$, $m-k$ dimensional. Fie $\boldsymbol{\alpha}'\mathbf{x}^2$ o combinaţie liniară de elementele lui $\mathbf{x}^2$ şi fie x_i un element al lui $\mathbf{x}'$.

DEFINIŢIA 9'. Maximul coeficientului de corelaţie dintre x_i şi combinaţia liniară $\boldsymbol{\alpha}'\mathbf{x}^2$ poartă numele de coeficientul de corelaţie multiplă dintre x_i şi vectorul $\mathbf{x}^2$ şi este notat prin $P_{i.\,k+1,\ldots,m}$.

$P_{i.\,k+1,\ldots,m}$ va depinde de $\boldsymbol{\Sigma}$. Pentru a pune în evidenţă această dependenţă demonstrăm

LEMA 2. *Cu $\boldsymbol{\alpha}$ arbitrar (nealeator), dispersia variabilei*

$$x_i - \boldsymbol{\alpha}'\mathbf{x}^2$$

îşi atinge minimul pentru

$$\boldsymbol{\alpha}' = \boldsymbol{\sigma}_{i.}\boldsymbol{\Sigma}_{22}^{-1} \tag{140}$$

unde $\boldsymbol{\sigma}_{i.}$ este linia i a matricii $\boldsymbol{\Sigma}_{12}$.

Demonstraţie. Cum

$$M(\mathbf{x}^1|\mathbf{x}^2) = \boldsymbol{\mu}^1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}^2 - \boldsymbol{\mu}^2) \tag{141}$$

avem

$$M(x_i|\mathbf{x}^{(2)}) = \mu_i + \boldsymbol{\sigma}_{i.}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}^2 - \boldsymbol{\mu}^2). \tag{141'}$$

Mai întîi vom arăta că $\mathbf{x}^2$ şi $x_i - \boldsymbol{\alpha}'\mathbf{x}^2$ sînt independente. Deoarece $x_i - \boldsymbol{\alpha}'\mathbf{x}^2$ şi $\mathbf{x}^2$ au o repartiţie comună normală, va trebui să arătăm că matricea lor de covarianţe este $\mathbf{0}$. Astfel

$$\begin{aligned} M[(x_i - \mu_i - \boldsymbol{\alpha}'(\mathbf{x}^2 - \boldsymbol{\mu}^2)'] &= M[(x_i - \mu_i)(\mathbf{x}^2 - \boldsymbol{\mu}^2)'] - \\ -\boldsymbol{\alpha}' M[(\mathbf{x}^2 - \boldsymbol{\mu}^2)(\mathbf{x}^2 - \boldsymbol{\mu}^2)'] &= \boldsymbol{\sigma}_{i.} - \boldsymbol{\alpha}'\boldsymbol{\Sigma}_{22}. \end{aligned} \tag{142}$$

Avem

$$\boldsymbol{\sigma}_{i.} - \boldsymbol{\alpha}'\boldsymbol{\Sigma}_{22} = \mathbf{0} \tag{143}$$

ceea ce stabileşte independenţa.

Fie $\boldsymbol{\gamma}$ orice vector de constante. Atunci

$$\begin{aligned} D^2(x_i - \boldsymbol{\gamma}\mathbf{x}^2) &= D^2[(x_i - \boldsymbol{\alpha}'\mathbf{x}^2) + (\boldsymbol{\alpha} - \boldsymbol{\gamma})'\mathbf{x}^2] = \\ &= D^2(x_i - \boldsymbol{\alpha}'\mathbf{x}^2) + (\boldsymbol{\alpha} - \boldsymbol{\gamma})'\boldsymbol{\Sigma}_{22}(\boldsymbol{\alpha} - \boldsymbol{\gamma}). \end{aligned} \tag{144}$$

Scanned with OKEN Scanner

Deoarece Σ_{22} este pozitiv definită, urmează că membrul stîng al relației (144) este minimizat cînd al doilea termen din membrul drept este zero. Însă acesta se realizează numai cînd

$$\gamma' = \alpha' = \sigma_{i.}\Sigma_{22}. \tag{145}$$

LEMA 3. *Fie* $\mathbf{x}$, $\boldsymbol{\mu}$, Σ, *partiționarea din lema 2 și combinația liniară* $\gamma'\mathbf{x}^2$, *cu* γ *nealeator. Atunci coeficientul de corelație dintre* x_i *și* $\gamma'\mathbf{x}^2$ *este maximizat pentru* (145).

Demonstrație. Din lema 2, pentru orice scalar c și vector γ, avem:

$$D^2(x_i - \alpha'\mathbf{x}^2) \leqslant D^2(x_i - c\gamma'\mathbf{x}^2). \tag{146}$$

Dezvoltînd ambii membrii, obținem

$$\sigma_{ii} - 2\sigma_{i.}\alpha + \alpha'\Sigma_{22}\alpha \leqslant \sigma_{ii} - 2c\sigma_{i.}\gamma + c^2\gamma'\Sigma_{22}\gamma. \tag{147}$$

Deoarece c este arbitrar, (147) are loc în particular pentru

$$c^2 = \frac{\alpha'\Sigma_{22}\alpha}{\gamma'\Sigma_{22}\gamma}. \tag{148}$$

Ținînd seama în (147) de (148), multiplicînd prin $\sigma_{ii}^{-1/2}$, găsim

$$\frac{\sigma_{i.}\alpha}{(\sigma_{ii}\alpha'\Sigma_{22}\alpha)^{1/2}} \geqslant \frac{\sigma_{i.}\gamma}{(\sigma_{ii}\gamma'\Sigma_{22}\gamma)^{1/2}}. \tag{149}$$

Însă membrul stîng al relației (149) este coeficientul de corelație dintre variabila aleatoare scalară x_i și $\alpha'\mathbf{x}^2$ și membrul drept este coeficientul de corelație dintre x_i și $\gamma'\mathbf{x}^2$ (pentru γ arbitrar).

Observație. Din lema 3 urmează că coeficientul de corelație multiplă dintre x_i și vectorul $\mathbf{x}^2$ este dat de

$$P_{i,k+1\ldots m} = \frac{\sigma_{i.}\alpha}{(\sigma_{ii}\alpha'\Sigma_{22}\alpha)^{1/2}} = \left(\frac{\sigma_{i.}\Sigma_{22}^{-1}\sigma_{i.}'}{\sigma_{ii}}\right)^{1/2}. \tag{150}$$

În cele ce urmează presupunem că $X_0, X_1, \ldots, X_k$, din definiția 9 au o repartiție comună normală multidimensională și că asupra acestuia avem o selecție de volum n.

DEFINIȚIA 10. Cantitatea $R_{0.1\ldots k}$ obținută înlocuind în (132) elementele matricei $\mathbf{V}$ prin estimatorii lor de verosimilitate maximă poartă numele de coeficientul de corelație multiplă de selecție.

$R_{0.12\ldots k}$ este o statistică și deci se pune problema determinării repartiției sale.

Cînd nu există pericolul de confuzie, indicii $0.12\ldots k$ se omit.

TEOREMA 11. (Hodgson (1967)) *Cu* $n > k + 1$, $R^2(1 - R^2)^{-1}$ *este repartizată ca*

$$[\chi^2(k-1) + \{U + P(1-P^2)^{-1/2}\chi(n-1)\}^2]/\chi^2(n-k-1) \tag{151}$$

unde variabilele χ^2 *și* U *sînt independente.*

Scanned with OKEN Scanner

Observînd că R nu poate fi negativ, din indentitatea

$$P(R \leqslant r) = P[R^2(1-R^2)^{-1} \leqslant r^2(1-r^2)^{-1}] =$$

$$= P[\chi^2(k-1) + \{U + P(1-P^2)^{-1/2}\chi(n-1)\}^2 -$$

(152)

$$-r^2(1-r^2)^{-1}\,\chi^2(n-k-1) < 0]$$

și aproximînd

a) raportul (151) sau

b) membrul stîng al ultimei inegalități din (152).

Hodgson sugerează că : (h fiind un număr pozitiv convenabil ales)

$$\frac{(n-k+h-2)^h[R^2(1-R^2)^{-1}]^h - \{k+2h-2+(n+h-2)\,P^2(1-P^2)^{-1}\}^h}{\sqrt{2h}[(n-k-1)^{2h-1}\{R^2(1-R^2)^{-1}\}^{2h}}$$

(152′)

$$+(2+P^2(1+P^2)^{-1})\{k+(n-1)P^2(1-P^2)^{-1}\}^{2h-1}]^{1/2}$$

are, aproximativ, o repartiție $N(0, 1)$;

(152″)

$$\{R^2(1-R^2)^{-1}\}^h$$

are, aproximativ, o repartiție normală cu valoarea medie

$$[(n-k+h-2)^{-1}\{k+2h-2+(n+h-2)[P^2(1-P^2)^{-1}]^h$$

și dispersia

$$2h^2(n-k-1)^{-(2h+1)}\{k+(n-1)P^2(1-P^2)^{-1}\}^{2h-1}(2n-k)(1-P^2)^{-1}.$$

Observații i) Repartiția lui R^2 a fost obținută inițial de Fisher (1928) folosind o metodă geometrică (de asemenea se poate vedea Soper (1929))

ii) Cazuri particulare au fost discutate anterior de Yule (1907) și Isserlis (1917).

Propoziția 38 (Fisher (1928)). *Densitatea de repartiție a statisticii* R^2 *este*

$$p_{R^2}(r^2) = \frac{\Gamma\left(\frac{1}{2}n\right)(1-P^2)^{(n-1)/2}}{\pi\Gamma\left(\frac{1}{2}(k-1)\right)\Gamma\left(\frac{1}{2}(n-k)\right)}(r^2)^{\frac{1}{2}k-1}\,(1-r^2)^{\frac{1}{2}(n-k)-1} \times$$

(153)

$$\times \int_0^\pi \int_{-\infty}^{\infty} \frac{\sin^{k-2}\theta}{(\operatorname{ch}\phi - P\cos\theta)^{n-1}}\,d\phi\,d\theta \quad (r^2 > 0).$$

Integrala poate fi evaluată dezvoltînd integrandul în puteri de $\cos\theta$ și integrînd termen cu termen. Rezultatul poate fi exprimat cu ajutorul

Scanned with OKEN Scanner

unei funcţii hypergeometrice

$$p_{R^2}(r^2) = \frac{(1-P^2)^{(n-1)/2}\,(r^2)^{(k-1)/2}\,(1-r^2)^{\frac{1}{2}(n-k-1)-1}}{B\left(\frac{1}{2}k, \frac{1}{2}(n-k-1)\right)} \times$$

(154)

$$\times F\left(\frac{n-1}{2}, \frac{n-1}{2}; \frac{k}{2}\,;\, P^2r^2\right), \quad (r^2>0).$$

Observaţii i) Excluzînd factorul $(1-P^2)^{(n-1)/2}$ şi funcţia hypergeometrică avem o repartiţie beta de parametrii $\frac{1}{2}k$, $\frac{1}{2}(n-k-1)$. Aceasta este repartiţia statistică R^2 cînd $P=0$.

ii) Din (154), sau prin dezvoltare directă este posibil să exprimăm $p_{R^2}(r^2)$ ca o mixtură de repartiţii beta cu parametrii $\left(\frac{1}{2}k+j\right), \frac{1}{2}(n-k-1)$, cu ponderile termeni în dezvoltarea lui

$$\left(\frac{1}{1-P^2} - \frac{P^2}{1-P^2}\right)^{-(n-1)/2}.$$

Are loc

TEOREMA 12 (Gurland (1968). *Avem*

(155)

$$P_{R^2}(r^2) =$$

$$= \sum_{j=0}^{\infty} b_j \left[B\left(\frac{1}{2}k+j, \frac{1}{2}(n-k-1)\right)\right]^{-1} (r^2)^{\frac{1}{2}k+j-1} (1-r^2)^{\frac{1}{2}(n-k-1)-1}, \quad (r>0)$$

cu

$$b_j = \frac{\Gamma\left(\frac{1}{2}(n-1)+j\right)}{j!\,\Gamma\left(\frac{1}{2}(n-1)\right)} \cdot \frac{(P^2)^j}{(1-P^2)^{(n-1)/2+j}}.$$

Gurland (1968) arată, de asemenea, că pentru $(n-k)$ impar,

$$R^2(1-P^2)(1-P^2R^2)^{-1}$$

este repartizată ca o mixtură de repartiţii beta cu parametrii $\left(\frac{1}{2}k+j\right)$; $\frac{1}{2}(n-k-1)$ şi ponderile date de termenii în dezvoltarea binomului

$$\{P^2 + (1-P^2)\}^{(n-k-1)/2}.$$

Mai precis, are loc

Scanned with OKEN Scanner

TEOREMA 13 (Gurland (1968)). *Avem*

$$P(R \leqslant r) = \sum_{j=0}^{(n-k-1)/2} C^j_{(n-k-1)} (P^2)^j \times$$

(156)

$$\times (1 - P^2)^{\frac{1}{2}(n-k-1)-j} I_{\frac{R^2(1-P^2)}{1-P^2R^2}} \left(\frac{1}{2}k + j, \frac{1}{2}(n-k-1)\right).$$

Observaţie În (156) sînt numai un număr finit de termeni.
Gurlan, Milton (1970) cercetează seriile (155) şi (156)
Repartiţia lui R este obţinută prin diferite metode de Garding (1941), Moran (1950), Soper (1929) şi Wilks (1932).

PROPOZIŢIA 39 (Banerjee (1952)). *Momentul de ordinul m faţă de zero al lui R este*

$$\mu'_m(R) = \frac{(1-P^2)^{(n-1)/2}\,\Gamma\left(\frac{1}{2}(k+m)\right)}{\Gamma\left(\frac{1}{2}(n-1+m)\right)} D'^m F\left(\frac{1}{2}(n-1+m), \frac{1}{2}n; \frac{1}{2}k; P^2\right)$$

(157)

unde D' reprezintă operatorul $\frac{1}{2}P^3\frac{\partial}{\partial P}$.

PROPOZIŢIA 40 (Wishart (1931)). *Valoarea medie şi dispersia statisticii R^2 sînt*

$$M(R^2) = 1 - \frac{n-k-1}{n-1}(1-P^2)F\left(1,1;\frac{1}{2}(n+1);P^2\right) =$$

(158)

$$= P^2 + \frac{k}{n-1}(1-P^2) - \frac{2(n-k-1)}{n^2-1}P^2(1-P^2) + \mathbf{0}(n^{-2})$$

şi respectiv

$$D^2(R^2) = \frac{(n-k)^2-1}{n^2-1}(1-P^2)^2 F\left(2,2;\frac{1}{2}(n+3);P^2\right) - [M(R^2)-1]^2 =$$

(159)

$$= \begin{cases} \dfrac{4P^2(1-P^2)^2(n-k-1)^2}{(n^2-1)(n+3)} + \mathbf{0}(n^{-2}), & P \neq 0 \\ \dfrac{2k(n-k-1)}{(n-1)^2(n+1)}, & P = 0. \end{cases}$$

Repartiţia statisticii R (sau a statisticii R^2) avînd o formă complicată, prezintă importanţă construcţia de aproximaţii. Este natural să folosim aproximaţia lui Fisher

$$z' = \frac{1}{2}\log\frac{1+R}{1-R} = \operatorname{arcth} R.$$

Scanned with OKEN Scanner

Totuși, ca și pentru corelația serială, această transformare nu este prea convenabilă.

PROPOZIȚIA 41 (Gajjar (1967)). *Statistica*

$$\sqrt{n-1}\,\operatorname{arcth} R \tag{160}$$

are pentru n tinzînd la infinit o repartiție χ *necentrată cu k grade de libertate și parametru de necentricitate*

$$(n-1)(\operatorname{arcth} P)^2.$$

Observație. Calcule numerice indică ca $\operatorname{arcth} R$ nu dă rezultate folositoare.

Khatri (1966) a propus două aproximații. Prima constă în a privi

$$k^{-1}(n-k-1)(1-P^2)\omega(P)\ R^2/(1-R^2)$$

cu

$$\omega(P) = [k + \{n-k-1+\sqrt{(n-1)(n-k-1)}\}P^2]\times$$

$$\times [k+(n-k-1)(2-P^2)P^2]^{-1}$$

ca fiind, aproximativ, repartizată F necentrată cu $(n-k-1)$ grade libertate și parametru de necentricitate

$$\frac{1}{2} P^2\omega(P)\sqrt{(n-1)(n-k-1)}.$$

A doua aproximație a lui Khatri folosește ca multiplicator al statisticii $R^2/(1-R^2)$ pe $(n-k-1)(1-P^2)\{(n-k-1)P^2+k\}^{-1}$ și repartiția, aproximativă, a acestei statistici este F centrat.

Khatri sugerează ca aceste aproximații să fie folosite pentru $n-k-1 \geqslant 100$ și cînd P^2 este mare o preferă pe a doua.

Hodgson (1967) construiește unele aproximații normale bazate pe reprezentarea (152) discutată anterior.

Aproximația propusă de Gurland (1968) (bazată de (155)) dă rezultate excelente. Aproximează repartiția statisticii $R^2(1-R^2)^{-1}$ prin aceea a statisticii $gG_{f,n-k}$ cu

$$g = \{n\theta(\theta+2)+k\}/(n\theta+k);\ f = (n\theta+k)^2/\{n\theta(\theta+2)+k\}$$

unde $\theta = P^2(1-P^2)^{-1}$ și $G_{f,n-k}$ este raportul a 2 variabile χ^2 independente.

Se poate arăta că $P(R \leqslant r|P)$ este o funcție descrescătoare de P. Rezolvînd ecuația $\int_{-\infty}^{R} p_R(r|P)dr = \alpha_1,\ 1-\alpha_2$ cu $(\alpha_1+\alpha_2=\alpha)$ se obțin limite inferioare și superioare pentru intervale de încredere pentru P avînd coeficientul de încredere $1-\alpha$.

Kramer (1963) dă tabele pentru obținerea limitelor inferioare de coeficient 0,95 (adică $\alpha_1 = 0{,}05,\ \alpha_2 = 0$).

Scanned with OKEN Scanner

§ 10. Repartiția normală tridimensională. Corelație parțială

DEFINIȚIA 11. Vom spune că vectorul aleator (X_1, X_2, X_3) urmează o repartiție normală tridimensională standardizată dacă are densitatea de repartiție

$$(161) \qquad P_{X_1, X_2, X_3}(x_1, x_2, x_3) = (2\pi)^{-3/2}\Delta^{-1/2}\exp\left[-\frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}A_{ij}x_ix_j\right]$$

unde:

$$\Delta = 1 - \rho_{23}^2 - \rho_{13}^2 - \rho_{12}^2 + 2\rho_{23}\rho_{13}\rho_{12},$$

$$A_{11} = (1 - \rho_{23}^2)\Delta^{-1},\ A_{22} = (1 - \rho_{13}^2)\Delta^{-1};\ A_{33} = (1 - \rho_{12}^2)\Delta^{-1};$$

$$A_{12} = A_{21} = (\rho_{13}\rho_{23} - \rho_{12})\Delta^{-1};\ A_{13} = A_{31} = (\rho_{12}\rho_{23} - \rho_{13})\Delta^{-1},$$

$$A_{23} = A_{32} = (\rho_{12}\rho_{13} - \rho_{23})\Delta^{-1}.$$

DEFINIȚIA 12. Dacă în (161) toți ρ_{ij}, $i, j = 1, 2, 3$ sînt zero și toți σ_i^2 sînt egali, vom spune că vectorul aleator (X_1, X_2, X_3) are o repartiție normală sferică.

Dacă în (161) toți ρ_{ij} sînt zero însă σ_i^2 nu sînt egali vom spune că vectorul aleator (X_1, X_2, X_3) are o repartiție normală elipsoidală.

PROPOZIȚIA 42. *Repartiția variabilei X_3, condiționată de X_1 și X_2, este normală de valoarea medie $\rho_{13.2}x_1 + \rho_{23.1}x_2$ și dispersie $(1 - \rho_{3.12}^2)$, unde: $\rho_{ij.k}$ reprezintă coeficientul de corelație dintre X_i și X_j dat X_k și $\rho_{3.12}^2$ este coeficientul de corelație multiplă al lui X_3, asupra lui X_1 și X_2.*

PROPOZIȚIA 43. *Repartiția comună a variabilelor X_1 și X_2 condiționate de X_3 este normală bidimensională cu repartiții marginale avînd mediile*

$$\rho_{13}x_3, \text{ respectiv } \rho_{23}x_3$$

și dispersiile

$$(1 - \rho_{13}^2), \text{ respectiv } (1 - \rho_{23}^2)$$

și cu coeficientul de corelație

$$(162) \qquad \rho_{12.3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{(1 - \rho_{13}^2)(1 - \rho_{23}^2)}}.$$

DEFINIȚIA 13. Parametrul $\rho_{12.3}$ poartă numele de coeficientul de corelație parțial între X_1 și X_2 dat X_3 (sau cu X_3 fixat).

Observații i) Cu toate că (162) a fost obținut în ipoteza de normalitate, vom defini coeficientul de corelație parțială prin (162) pentru orice repartiție.

ii) (162) este simetric în primii doi indici; al treilea indice se referă la variabila fixată.

Scanned with OKEN Scanner

PROPOZIȚIA 44 (Nabeya (1952)) $\nu_{rst} = M(|X_1^r X_2^s X_3^t|)$ *are următoarele valori*

$$\nu_{111} = (2/\pi)^{3/2}(\Delta^{1/2} + \sum{}^*(\rho_{23} + \rho_{12}\rho_{13}) \arcsin \rho_{23.1}) \tag{163}$$

unde $\sum^*$ *notează suma ciclică*

$$\nu_{211} = 2\pi^{-1}[(\rho_{23} + 2\rho_{12}\rho_{13})\arcsin \rho_{23} + (1 + \rho_{12}^2 + \rho_{13}^2)\sqrt{1 - \rho_{23}^2}]. \tag{164}$$

PROPOZIȚIA 45 (Kamat) (1958)). *Pentru momentele incomplete:*

$$[r, s, t] = (2\pi)^{-3/2}\Delta^{-1/2}\int_0^\infty\int_0^\infty\int_0^\infty x_1^r x_2^s x_3^t \exp\left[-\frac{1}{2}\sum_{i=1}^3\sum_{j=1}^3 A_{ij}x_i x_j\right]dx_1\,dx_2\,dx_3$$

(165)
avem

$$[1, 0, 0] = (2\pi)^{-3/2}\left\{\frac{1}{2}\pi + \arcsin \rho_{23.1} + \rho_{12}\left(\frac{1}{2}\pi + \arcsin \rho_{13.2}\right) + \right.$$

$$\left. + \rho_{13}\left(\frac{1}{2}\pi + \arcsin \rho_{12.3}\right)\right\};$$

$$[2, 0, 0] = (4\pi)^{-1}\left\{\frac{1}{2}\pi + \sum_{i<j}^3 \arcsin \rho_{ij} + \Delta\rho_{23}\sqrt{1 - \rho_{23}^2} + (2\rho_{12}\rho_{13} - \rho_{23})\times\right.$$

$$\left.\times\sqrt{1 - \rho_{23}^2} + \rho_{12}\sqrt{1 - \rho_{12}^2} + \rho_{13}\sqrt{1 - \rho_{13}^2}\right\};$$

(166)

$$[1, 1, 0] = (4\pi)^{-1}\left\{\rho_{12}\left(\frac{1}{2}\pi + \sum_{i<j}^3 \arcsin \rho_{ij}\right) + \right.$$

$$\left. + \sqrt{1 - \rho_{12}^2} + \rho_{13}\sqrt{1 - \rho_{13}^2} + \rho_{23}\sqrt{1 - \rho_{23}^2}\right\};$$

$$[1,1,1] = (2\pi)^{-3/2}\left[\Delta^{1/2} + \sum{}^*(\rho_{23} + \rho_{12}\rho_{13})\left(\frac{1}{2}\pi + \arcsin \rho_{23.1}\right)\right].$$

Kamat (1965) dă formule pentru $[r, s, t]$ pentru toți r, s, t cu $r + s + t \leqslant 3$.

Alte formule sînt date de Haldane (1942).

Considerăm repartiția normală tridimensională standardizată de vector valoare medie (0, 0, 0) și matrice de covariație:

$$\mathbf{C} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}$$

Scanned with OKEN Scanner

Analog funcţiilor $\Phi(h,\ k;\ \rho)$ şi $L(h,\ k;\ \rho)$ (din § 1), avem:

(167) $\Phi(h_1,\ h_2,\ h_3;\ \rho_{23},\ \rho_{13},\ \rho_{12}) = P[(X_1 < h_1) \cap (X_2 < h_2) \cap (X_3 < h_3)]$

şi

(168) $L(h_1,\ h_2,\ h_3;\ \rho_{23},\ \rho_{13},\ \rho_{12}) = P[(X_1 > h_1) \cap (X_2 > h_2) \cap (X_3 > h_3)]$.

Ecuaţia (27) se generalizează la

(169) $$\Phi(0,\ 0,\ 0;\ \rho_{23},\ \rho_{13},\ \rho_{12}) = L(0,\ 0,\ 0;\ \rho_{23},\ \rho_{13},\ \rho_{12})$$

Ruben (1954) dă o tabelă pentru $\frac{1}{2} - \frac{3}{4}\pi \arccos\rho$ (ce corespunde la $\rho_{23} = \rho_{13} = \rho_{12} = \rho$) cu 8 zecimale pentru $\rho^{-1} = 2(1)\ 11$ (în notaţiile sale $\bar{V}_{n,n}(x) = \bar{u}_n(x)$ pentru $n = 3$). Pentru $\Phi\left(h,\ h,\ h;\ \frac{1}{2},\ \frac{1}{2},\ \frac{1}{2}\right)$ au dat tabele:

i) Teichroew (1955) cu 5 zecimale pentru $h\sqrt{2} = 0(0{,}01)\ 6{,}09$ şi

ii) Somerville (1954) cu 4 zecimale pentru $h = 0(0{,}1)\ 2{,}0\ (0{,}5)\ 3{,}0$.

Steck (1958) exprimă Φ cu ajutorul funcţiei $S(h,\ a,\ b)$ definite (pentru $h,\ a,\ b > 0$) ca

(170) $$\frac{1}{4\pi}\operatorname{arctg}\frac{b}{\sqrt{(1+a^2+a^2b^2)}} + P[(0 < U_1 < U_2 + bU_3) \cap (0 < U_2 < h) \cap (U_3 > aU_2)],$$

unde U_1, U_2 şi U_3 sînt variabile independente repartizate $N(0,1)$. Pentru valori negative ale lui h, a sau b, $S(h,\ a,\ b)$ este definită de:

(171) $$S(h,\ -a,b) = S(h,\ a,\ b)$$

(172) $$S(h,\ a,\ -b) = -S(h,\ a,\ b)$$

(173) $$S(-h,\ a,\ b) = \frac{1}{2\pi}\operatorname{arctg}\frac{b}{\sqrt{1+a^2+a^2b^2}} - S(h,\ a,\ b).$$

De asemenea:

(174) $$S(0,\ a,\ b) = \frac{1}{2}S(\infty,\ a,\ b) = \frac{1}{4\pi}\operatorname{arctg}\frac{b}{\sqrt{1+a^2+a^2b^2}};$$

(175) $$S(h,\ 0,\ b) = \frac{1}{2\pi}\Phi(h)\operatorname{arctg} b$$

(176) $$S(h,\ a,\ 0) = S(h,\ \infty,\ b) = 0$$

(177) $$S(h,\ a,\ \infty) = \begin{cases} \frac{1}{2}\left[\frac{1}{2}\Phi(h) + T(h, |a|)\right] - \frac{1}{2\pi}\operatorname{arctg}|a|, & h \geqslant 0 \\ \frac{1}{2}\left[\frac{1}{2}\Phi(h) - T(h, |a|)\right], & h < 0 \end{cases}$$

Scanned with OKEN Scanner

unde

$$T(h,\lambda)=\frac{1}{2\pi}\int_h^\infty \varphi(x_1)\int_{\lambda x_1}^\infty \varphi(x_2)\mathrm{d}x_2\mathrm{d}x_1 = \tag{178}$$

$$=\frac{1}{2\pi}\int_0^\lambda (1+x^2)^{-1}\exp\left[-\frac{1}{2}h^2(1+x^2)\right]\mathrm{d}x =$$

$$=\frac{1}{2\pi}\left[\operatorname{arc\,tg}\,\lambda-\sum_{j=0}^{\infty}C_j\lambda^{2j+1}\right] \quad \text{(Owen 1956))}$$

cu

$$c_j=\frac{(-1)^j}{2j+1}\left[1-\mathrm{e}^{-(1/2)h^2}\sum_{i=0}^{j}\frac{\left(\frac{1}{2}h^2\right)^i}{i!}\right].$$

Owen (1956) a dat valorile funcției $T(h,\lambda)$ cu 6 zecimale pentru $h=0(0{,}01)2{,}00(0{,}02)3{,}00$; $\lambda=0{,}25(0{,}25)1{,}00$ și pentru $h=0(0{,}05)4{,}75$; $\lambda=0(0{,}25)1{,}00$; ∞ (Owen (1957)).

Formulele pentru $\Phi(h_1,h_2,h_3;\rho_{23},\rho_{13},\rho_{12})$ sînt

i) $h_1,h_2,h_3\geqslant 0$ (sau $h_1,h_2,h_3\leqslant 0$)

$$\Phi(h_1,h_2,h_3;\rho_{23},\rho_{13},\rho_{12})=\sum_{j=1}^{3}\left(1-\frac{1}{2}\delta_{a_jc_j}\right)\Phi(h_j)+\frac{1}{4}\sum_{i<j}^{3}\sum^{3}\delta_{h_ih_j}+$$

$$+\frac{1}{2}\left[\sum_{i<j}^{3}\sum^{3}L(h_i,h_j;\rho_{ij})-3\right]- \tag{179}$$

$$-\sum_{j=1}^{3}[S(h_j,a_j,b_j)+S(h_j,c_j,d_j)],$$

unde

$$a_1=\frac{h_2-h_1\rho_{12}}{h_1\sqrt{1-\rho_{12}^2}};\quad a_2=\frac{h_3-h_2\rho_{23}}{h_2\sqrt{1-\rho_{23}^2}};\quad a_3=\frac{h_1-h_3\rho_{13}}{h_3\sqrt{1-\rho_{13}^2}};$$

$$c_1=\frac{h_3-h_1\rho_{13}}{h_1\sqrt{1-\rho_{13}^2}};\quad c_2=\frac{h_1-h_2\rho_{12}}{h_2\sqrt{1-\rho_{12}^2}};\quad c_3=\frac{h_2-h_3\rho_{23}}{h_3\sqrt{1-\rho_{23}^2}};$$

$$b_1=\sqrt{(1-\rho_{12}^2)(1-\rho_{13}^2)}\,(c_1a_1^{-1}-\rho_{22\cdot1})\Delta^{-1/2};$$

$$d_1=\sqrt{(1-\rho_{12}^2)(1-\rho_{13}^2)}\,(a_1c_1^{-1}-\rho_{23\cdot1})\Delta^{-1/2}$$

și așa mai departe; și δ_{hk} definite ca în propoziția 14

ii) $h_1,h_2\geqslant 0$; $h_3<0$ (sau $h_1,h_2\leqslant 0$, $h_3>0$)

$$\Phi(h_1,h_2,h_3;\rho_{23},\rho_{13},\rho_{12})=L(h_1,h_2;\rho_{12})+\Phi(h_1)+\Phi(h_2)-$$

$$-1-\Phi(h_1,h_2,-h_3;-\rho_{23},\rho_{13},\rho_{12}). \tag{180}$$

Scanned with OKEN Scanner

Steck (1958) dă tabele pentru funcţia $S(h, a, b)$ cu 7 zecimale pentru $a = 0, 0(0,1)\ 2,0(0,2)5,0(0,5)8,0$; $b = 0,1(0,1)1,0$ şi h crescînd de la 0 prin intervale de 0,1 la o limită astfel încît:

$$S(h, a, b) \simeq \frac{1}{2\pi} \Phi\left(h \sqrt{\left(1 + a^2 + \frac{1}{4} a^2 b_2\right)}\right) \operatorname{arc\,tg}\left(\frac{b}{\sqrt{(1 + a^2 + a^2 b^2)}}\right) \tag{181}$$

cu o eroare mai mică decît 5×10^{-5}. Pentru valori ale lui $b > 1$, pot fi folosite formulele:

$$S(h, a, b) = \left[\Phi(h) - \frac{1}{2}\right] T(ah, b) - \left[\Phi(hab) - \frac{1}{2}\right] T(ah, a^{-1}) + \tag{182}$$

$$+ S(hab, b^{-1}, a^{-1}) \text{ (pentru } a > 1)$$

sau

$$S(h, a, b) = \frac{1}{4} \Phi(h) + \left[\Phi(hab) - \frac{1}{2}\right] T(h, a) - S(hab, (ab)^{-1}, a) \tag{183}$$

$$- S(h, ab, b^{-1}) \text{ (pentru } a < 1).$$

Estimatorul de verosimilitate maximă al coeficientului de corelaţie multiplă al variabilei X_1 cu X_2 şi X_3

$$R^2 = (\rho_{12}^2 + \rho_{13}^2 - 2\rho_{12}\rho_{13}\rho_{23})/(1 - \rho_{23}^2) \tag{184}$$

este

$$\hat{R}^2 = (\hat{\rho}_{12}^2 + \hat{\rho}_{13}^2 - 2\hat{\rho}_{12}\hat{\rho}_{13}\hat{\rho}_{23})/(1 - \hat{\rho}_{23}^2). \tag{185}$$

Reluăm problema definirii corelaţiei parţiale în cazul multidimensional.

Considerăm cazul particular al repartiţiei normale p-dimensionale şi să fixăm $(p - 2)$ dintre componentele acesteia. Atunci coeficientul de corelaţie parţială al celorlalte două (să presupunem X_1, X_2)

$$\rho_{12 \cdot 34 \cdots p} = \frac{-C_{12}}{(C_{11} C_{22})^{1/2}} \tag{186}$$

unde C_{ij} este cofactorul lui ρ_{ij} în

$$|C| = \begin{vmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1p} \\ & 1 & \rho_{23} & \cdots & \rho_{2p} \\ & & 1 & \cdots & \rho_{3p} \\ & & & \ddots & \vdots \\ & & & & 1 \end{vmatrix} \tag{187}$$

Observație. (186) reprezintă definiția generală a coeficientului de corelație parțială dintre X_1 și X_2 cu $X_3, \ldots, X_p$ fixate.

Considerăm problema din alt punct de vedere. Fie

$$f(x_1, \ldots, x_k) | x_{k+1}, \ldots, x_p)$$

densitatea de repartiție comună a variabilelor $X_1, \ldots, X_k$ cînd $X_{k+1}, \ldots, X_p$ sînt fixate și

$$g(x_{k+1}, \ldots, x_p)$$

densitatea de repartiție marginală a variabilelor $X_{k+1}, \ldots, X_p$.

Funcția caracteristică a variabilelor $X_1, \ldots, X_p$ este

$$\phi(t_1, \ldots, t_p) =$$

$$= \int \ldots \int f(x_1, \ldots, x_k | x_{k+1}, \ldots, x_p) g(x_{k+1}, \ldots, x_p) \exp\left(\sum_1^p i t_j x_j \right) dx_1 \ldots dx_p =$$

$$= \int \ldots \int \phi_k(t_1, \ldots, t_k | x_{k+1}, \ldots, x_p) g(x_{k+1}, \ldots, x_p) \exp\left(\sum_{j=k+1}^p i t_j x_j \right) dx_{k+1} \ldots dx_p$$

unde $\phi_k(t_1, \ldots, t_k | x_{k+1}, \ldots, x_p)$ este funcția caracteristică a variabilelor $X_1, \ldots, X_k$ condiționată de $X_{k+1}, \ldots, X_n$. Din teorema de inversiune multidimensională urmează că

$$(188)\quad \phi_k g = \frac{1}{(2\pi)^{p-k}} \int \ldots \int \phi(t_1, \ldots, t_p) \exp\left(- \sum_{j=k+1}^p i t_j x_j \right) dt_{k+1} \ldots dt_p.$$

Dacă în (188) punem $t_1 = t_2 = \ldots = t_k = 0$ obținem

$$(189)\quad g = \frac{1}{(2\pi)^{p-k}} \int \ldots \int \phi(0, \ldots, 0, t_{k+1}, \ldots, t_p) \exp\left(- \sum_{j=k+1}^p i t_j x_j \right) dt_{k+1} \ldots dt_p$$

(deoarece în acest caz $\phi_k = 1$)

Deci împărțind (188) prin (189), găsim

$$(190)\quad \phi_k = \frac{\int \ldots \int \phi(t_1, \ldots, t_p) \exp\left(- \sum_{j=k+1}^p i t_j x_j \right) dt_{k+1} \ldots dt_p}{\int \ldots \int \phi(0, \ldots, 0, t_{k+1}, \ldots, t_p) \exp\left(- \sum_{j=k+1}^p i t_j x_j \right) dt_{k+1} \ldots dt_p}.$$

Scanned with OKEN Scanner

Dacă acum presupunem că $X_1, \ldots, X_p$ au repartiție normală p-dimensională, numărătorul din (190) devine

$$\exp\left(-\frac{1}{2}\sum_{i,j=1}^{p}\rho_{ij}t_it_j - \sum_{j=k+1}^{p} it_jx_j\right) =$$

$$= \exp\left(-\frac{1}{2}\sum_{i,j=1}^{k}\rho_{ij}t_it_j\right)\exp\left(-\frac{1}{2}\sum_{i,j=k+1}^{p}\rho_{ij}t_it_j\right)\times$$

$$\times \exp\left(-\sum_{i=1}^{k}\sum_{j=k+1}^{p}\rho_{ij}t_it_j\right)\exp\left(-\sum_{j=k+1}^{p} it_jx_j\right) = \tag{191}$$

$$= \exp\left(-\frac{1}{2}\sum_{i,j=1}^{k}\rho_{ij}t_it_j\right)\exp\left(-\frac{1}{2}\sum_{i,j=k+1}^{p}\rho_{ij}t_it_j\right)\times$$

$$\times \exp\left\{-\sum_{j=k+1}^{p} it_j\left(x_j - i\sum_{i=1}^{k}\rho_{ij}t_i\right)\right\}.$$

Acum integrala în raport cu $t_{k+1}, \ldots, t_p$ a ultimilor doi factori din membrul drept ai relației (191) este rezultatul inversării funcției caracteristice a repartiției normale a variabilelor $X_{k+1}, \ldots, X_p$ cu X_j măsurat de la valoarea $i\sum_{m=1}^{k}\rho_{jm}t_m$. Această schimbare a originii nu afectează corelațiile. Dacă notăm prin **D** matricea de corelații a variabilelor $X_{k+1}, \ldots, X_p$ obținem pentru integrala din (191)

$$\exp\left(-\frac{1}{2}\sum_{i,j=1}^{k}\rho_{ij}t_it_j\right)\times$$

$$\times \exp\left\{-\frac{1}{2}\sum_{i,j=k+1}^{p}D^{ij}\left(x_i - i\sum_{m=1}^{k}\rho_{im}t_m\right)\left(x_j - i\sum_{m=1}^{k}\rho_{jm}t_m\right)\right\}.$$

Din (190) găsim că

$$\phi_k(t_1, \ldots, t_k | x_{k+1}, \ldots, x_p) = \exp\left(-\frac{1}{2}\sum_{i,j=1}^{k}\rho_{ij}t_it_j\right)\times$$

$$\times \exp\left\{-\frac{1}{2}\sum_{i,j=k+1}^{p}D^{ij}\left(x_i - i\sum_{m=1}^{k}\rho_{im}t_m\right)\left(x_j - i\sum_{m=1}^{k}\rho_{jm}t_m\right) + \right. \tag{192}$$

$$\left. + \frac{1}{2}\sum_{i,j=k+1}^{p}D^{ij}x_ix_j\right\}.$$

Astfel dacă σ'_{uv} reprezintă corelația variabilelor X_u și X_v în repartiția condiționată a variabilelor $X_1, \ldots, X_k$ și σ_{uv} covarianța lor necondiționată,

prin identificarea coeficienţilor lui $t_u\, t_v$ în (192) găsim

$$\sigma'_{uv} = \sigma_{uv} - \sum_{l,j=k+1}^{p} D^{lj} \rho_{lu} \rho_{jv}. \tag{193}$$

În (193) avem variabilele iniţiale standardizate. Dacă acum ne reîntoarcem la variabilele iniţiale (nestandardizate), $D^2(X_i) = \sigma_i^2$, fiecare ρ fiind înlocuit prin σ corespunzător, D^{ij} este înlocuit prin $D^{ij}/(\sigma_i \sigma_j)$ şi avem forma mai generală a lui (193)

$$\sigma'_{uv} = \sigma_{uv} - \sum_{l,j=k+1}^{p} D^{lj} \sigma_{lu} \sigma_{jv} / (\sigma_l \sigma_j). \tag{194}$$

(194) nu depinde de valorile în care $X_{k+1}, \ldots, X_p$ sînt fixate.

Observaţie. Dacă fixăm numai o variabilă, de exemplu X_p avem $D^{pp} = 1$ şi covarianţa condiţionată (194) devine

$$\sigma'_{uv} = \sigma_{uv} - \sigma_{pu} \sigma_{pv} / \sigma_p^2 = \sigma_u \sigma_v (\rho_{uv} - \rho_{up} \rho_{vp}), \tag{195}$$

şi dacă $u = v$, din (195) avem dispersia condiţionată a variabilei u

$$\sigma'^2_u = \sigma_u^2 (1 - \rho^2_{up}).$$

Din ultimele două formule obţinem coeficientul de corelaţie condiţionat

$$\rho_{uv.p} = \frac{\rho_{uv} - \rho_{up} \rho_{vp}}{\{(1 - \rho_{up}^2)(1 - \rho_{vp}^2)\}^{1/2}}, \tag{162'}$$

o altă formă a lui (162).

Dacă fixăm variabilele cu excepţia lui X_1 şi X_2, din (194) avem:

$$\rho_{12.34\ldots p} = \frac{\rho_{12} - \sum_{l,j=3}^{p} D^{lj} \rho_{l1} \rho_{j2}}{\left\{\left(1 - \sum_{l,j=3}^{p} D^{lj} \rho_{l1} \rho_{j1}\right)\left(1 - \sum_{l,j=3}^{p} D^{lj} \rho_{l2} \rho_{j2}\right)\right\}^{1/2}}. \tag{196}$$

Din (187) obţinem că minorul lui ρ_{12}

$$\begin{vmatrix} \rho_{21} & \rho_{23} & \rho_{24} \cdots \rho_{2p} \\ \rho_{31} & 1 & \rho_{34} \cdots \rho_{3p} \\ \vdots & \vdots & \vdots \quad \vdots \\ \rho_{p1} & \rho_{p3} & \rho_{p4} \cdots 1 \end{vmatrix} = \begin{vmatrix} \rho_{21} & \rho_{32} \cdots \rho_{p2} \\ \rho_{31} & \\ \vdots & \mathbf{D} \\ \rho_{p1} & \end{vmatrix}$$

poate fi dezvoltat după prima linie şi coloană ca

$$\rho_{21} |\mathbf{D}| - \sum_{l,j=3}^{p} D^{lj} \rho_{l1} \rho_{j2},$$

Scanned with OKEN Scanner

Scanned with OKEN Scanner

și similar pentru minorii lui ρ_{11}, ρ_{22}. Astfel (196) poate fi scris ca

$$\rho^2_{12.34\ldots p} = \frac{C^2_{12}}{C_{11}C_{22}} \tag{196'}$$

care este din nou (186).

§ 11. Variabile dichotomizate

11.1. Corelație tetrachorică

Cînd atît X_1 cît și X_2 sînt dichotomizate, datele pot fi prezentate în forma unei tabele 2×2. Simbolurile a, b, c și d reprezintă frecvențele observate, într-o selecție de volum n, ale evenimentelor

	$X_1 \leqslant x_{10}$	$X_1 > x_{10}$
$X_2 > x_{20}$	c	d
$X_2 \leqslant x_{20}$	a	b

, $a+b+c+d = n$

a: $(X_1 \leqslant x_{10}) \cap (X_2 \leqslant x_{20})$

b: $(X_1 > x_{10}) \cap (X_2 \leqslant x_{20})$

c: $(X_1 \leqslant x_{10}) \cap (X_2 > x_{20})$

d: $(X_1 > x_{10}) \cap (X_2 > x_{20})$.

Cu toate că există numai 3 observații distincte și cel puțin 5 parametrii necunoscuți (7 dacă x_{10} și x_{20} sînt necunoscute) este posibil să construim un estimator pentru ρ. Metoda pe care o vom prezenta a fost propusă de Pearson (1901)

Frecvența observată relativă a evenimentului $X_1 \leqslant x_{10}$ este $(a + c)/n$ și estimăm (în ipoteză că această dublă dichotomizare se realizează asupra componentelor unui vector normal bidimensional)

$$h_1 = (x_{10} - \mu_1)/\sigma_1 \text{ prin } \tilde{h}_1$$

$\tilde{h}_1$ satisfăcînd ecuația

$$\Phi(\tilde{h}_1) = \frac{a + c}{n}. \tag{197}$$

Similar:

$$h_2 = (x_{20} - \mu_2)/\sigma_2 \text{ este estimat prin } \tilde{h}_2,$$

$\tilde{h}_2$ satisfăcînd ecuaţia

$$\Phi(\tilde{h}_2) = (a + b)/n. \tag{198}$$

Atunci ρ este estimat prin $\tilde{\rho}$, unde

$$L(\tilde{h}_1, \tilde{h}_2; \tilde{\rho}) = d/n \text{ } (L \text{ fiind dată de (24)}). \tag{199}$$

DEFINIŢIA 14. $\tilde{\rho}$ astfel obţinut poartă numele de coeficient de corelaţie tetrachoric.

Observaţie. Denumirea de corelaţie tetrachorică provine din faptul că este bazat pe o tabelă tetrachorică (cu 4 intrări).

PROPOZIŢIA 46. (Hamdan (1970)) $\tilde{\rho}$ *este echivalent cu estimatorul de verosimilitate maximă pentru* ρ *în tabele* 2×2.

Ecuaţia (199) poate fi rezolvată printr-un proces iterativ, folosind tabelele cu funcţia $L(h_1, h_2; \rho)$

Înainte de a avea aceste tabele, s-au folosit metode analitice aproximative.

PROPOZIŢIA 47. (Pearson (1901)) *Are loc dezvoltrea*

$$L(h_1, h_2; \rho) =$$

$$= \Phi(h_1)\Phi(h_2) + \varphi(h_1)\varphi(h_2)\left[\rho + \frac{\rho^2}{2!} h_1 h_2 + \frac{\rho^3}{3!}(h_1^2 - 1)(h_2^2 - 1) + \dots\right] =$$

$$= \sum_{j=0}^{\infty} \tau_j(h_1)\tau_j(h_2)\rho^j, \tag{200}$$

unde

$$\tau_j(h) = \frac{(-1)^{j-1}}{\sqrt{j!}} \frac{\mathrm{d}^{j-1}\varphi(h)}{\mathrm{d}h^{j-1}} \quad (j \geqslant 1)$$

şi

$$\tau_0(h) = \Phi(h).$$

DEFINIŢIA 15. $\tau_j(h)$ poartă numele de funcţia tetrachorică de ordin j.

Lee (1917) tabelează $\tau_j(h)$ cu 7 zecimale pentru $j = 0(1)19$; $h = 0{,}0(0{,}1)4{,}0$. Aceste tabele sînt incluse în Tables for Statisticians and Biometricians (Pearson (1931)).

O altă dezvoltare este

$$L(h_1, h_2; \rho) = \Phi(h_1)\,\Phi(h_2) +$$

$$+ \varphi(h_1)\varphi(h_2)\left[\theta + \frac{\theta^2}{2} h_1 h_2 + \frac{\theta^3}{3!}(h_1^2 + h_2^2 - h_1^2 h_2^2) + \right.$$

$$\left. + \frac{\theta^4}{4!} h_1 h_2 \{5 - 3(h_1^2 + h_2^2) + h_1^2 h_2^2\} + \dots\right] \tag{201}$$

cu $\theta = \arcsin \rho$.

Scanned with OKEN Scanner

PROPOZIȚIA 48. (Pearson (1901)). *Avem*

$$R_1 = \frac{ad - bc}{n^2\varphi(\tilde{h}_1)\varphi(\tilde{h}_2)} = \tilde{\rho} + \frac{\tilde{\rho}^2}{2!}\tilde{h}_1\tilde{h}_2 + \tag{202}$$

$$+\frac{\tilde{\rho}^3}{3!}(\tilde{h}_1^2 - 1)(\tilde{h}_2^2 - 1) + \ldots + \frac{\tilde{\rho}^j}{j!}H_{j-1}(\tilde{h}_1)H_{j-1}(\tilde{h}_2) + \ldots$$

Rezultă ținînd seama de valorile funcțiilor $\Phi(\tilde{h}_1)$ și $\Phi(\tilde{h}_2)$ din (197) și (198) în (200).
Urmează că pentru valori mici ale lui $\tilde{\rho}$,

$$\tilde{\rho} \simeq R_1.$$

Anterior, Yule (1897) sugerează estimatorul

$$R_2 = (ad - bc)/(ad + bc). \tag{203}$$

Pearson (1901) consideră de asemenea estimatorii

$$R_3 = \cos\,[\pi(1 + ad/bc)^{-1}], \tag{204}$$

$$R_4 = \sin\left[(\pi/2)\left\{1 + \frac{2bc}{ad - bc}\cdot\frac{n}{b + c}\right\}^{-1}\right], \quad (ad \geqslant bc) \tag{205}$$

și

$$R_5 = \sin\left[(\pi/2)\left\{1 + \frac{4abcdn^2}{(ad - bc)^2(a + d)(b + c)}\right\}^{-1/2}\right], \tag{206}$$

și face comparații între ei.

Rezultatele lui Pearson (1901) sînt extinse de Castellan (1966) care găsește că un estimator propus de Camp (1931) dă rezultate apropiate de cele ale lui $\hat{\rho}$. Acest estimator este construit după cum urmează :

Mai întîi aranjăm (dacă este necesar schimbînd semnele) ca

$$a + c \geqslant b + d.$$

Calculăm $\tilde{h}_2^-$, $\tilde{h}_2^+$ din

$$\Phi(\tilde{h}_2^-) = a/(a + c)\,;\ \Phi(\tilde{h}_2^+) = b/(b + d) \tag{207}$$

și

$$M = \frac{(a + c)(b + d)}{n^2}\cdot\frac{\tilde{h}_2^- + \tilde{h}_2^+}{\varphi(\tilde{h}_1)}.$$

Atunci estimatorul este

$$R_6 = M(1 + \theta^2 M)^{-1}, \tag{208}$$

Scanned with OKEN Scanner

unde θ poate fi găsit prin interpolare în următoarea tabelă,

$$(a+c)/n = 0{,}50\ \ 0{,}55\ \ 0{,}60\ \ 0{,}65\ \ 0{,}70\ \ 0{,}75\ \ 0{,}80\ \ 0{,}85\ \ 0{,}90$$

$$\theta = 0{,}64\ \ 0{,}63\ \ 0{,}63\ \ 0{,}63\ \ 0{,}62\ \ 0{,}61\ \ 0{,}60\ \ 0{,}58\ \ 0{,}56$$

Pentru multe scopuri este suficient să luăm $\theta = \frac{5}{8}$. În general se obține o valoare diferită pentru R_6, dacă X_1 și X_2 sînt interschimbabile.

Abaterea medie pătratică a lui $\hat{\rho}$, pentru valori mari ale lui n, este aproximativ (Pearson (1901)).

$$\frac{1}{\sqrt{n}} \frac{\sqrt{\Phi(h_1)[1-\Phi(h_2)]\Phi(h_2)[1-\Phi(h_2)]}}{\varphi(h_1)\varphi(h_2)}. \tag{209}$$

Adeseori (209) este estimată prin

$$\frac{1}{n^{5/2}} \cdot \frac{\sqrt{(a+b)(c+d)(a+c)(b+d)}}{\varphi(\hat{h}_1)\varphi(\hat{h}_2)} \tag{210}$$

11.2. Corelația biserială

Dacă numai una dintre cele două variabile, să presupunem X_2, este dichotomizată, putem privi datele ca reprezentînd n perechi de variabile aleatoare (X_{1j}, Y_j) unde:

$$Y_j = \begin{cases} 1, & X_{2j} > x_{20} \\ 0, & X_{2j} \leqslant x_{20}. \end{cases}$$

Coeficientul de corelație dintre X_{1j} și Y_j, să-l notăm prin ρ', este legat de ρ prin formula

$$\rho' = \rho\varphi(h_2)[\Phi(h_2)\{1-\Phi(h_2)\}]^{-1/2} \tag{211}$$

unde $h_2 = (x_{20} - \mu_2)/\sigma_2$.

Luăm ca estimator al lui ρ, statistica $\rho^* = (\textit{coeficientul de corelație dintre } X_1 \textit{ și } Y) \cdot \frac{\sqrt{\Phi(\hat{h}_2)[1-\Phi(\hat{h}_2)]}}{\varphi(\hat{h}_2)}$, unde $\Phi(\hat{h}_2) = \bar{Y} = n^{-1}\sum_{j=1}^{n} Y_j$. Deoarece $\sum_{j=1}^{n}(Y_j - \bar{Y})^2 = n\Phi(\hat{h}_2)[1-\Phi(\hat{h}_2)]$, urmează că

$$\rho^* = \frac{n^{-1}\sum_{j=1}^{n}(X_{1j}-\bar{X}_1)(Y_j-\bar{Y})}{\varphi(\hat{h}_2)\left[n^{-1}\sum_{j=1}^{n}(X_{1j}-\bar{X}_1)^2\right]^{1/2}}. \tag{212}$$

Scanned with OKEN Scanner

DEFINIȚIA 16. (Pearson (1903)). ρ^* poartă numele de coeficientul de corelație biserial.

PROPOZIȚIA 49. (Soper (1915)). *Pentru valori mari ale lui* n

$$nD^2(\rho^*) \simeq \rho^4 + (h_2^2\rho^2 - 1)\frac{\Phi(h_2)[1-\Phi(h_2)]}{[\varphi(h_2)]^2} + \rho^2\left\{\frac{(2\Phi(h_2)-1)h_2}{\varphi(h_2)} - \frac{5}{2}\right\}. \tag{213}$$

Tate (1955) dă valorile rădăcinii pătrate a membrului drept al relației (213) cu 3 zecimale pentru $\rho = 0{,}0(0{,}1)1{,}0$; $\Phi(h_2) = 0{,}05(0{,}05)0{,}50$, valoarea fiind neschimbată dacă semnul lui ρ sau h_2 este schimbat.

PROPOZIȚIA 50. (Tate (1955)). i) *Estimatorii de V.M. pentru* μ_1 *și* σ_1^2 *sînt* $\bar{X}_1$ *și respectiv* $n^{-1}\sum_{j=1}^{n}(X_{1j} - \bar{X}_1)^2 = S^2$

ii) $\hat{\rho}$ și $\hat{h}_2$, *estimatorii de V.M. pentru* ρ *și respectiv* h_2 *satisfac ecuațiile:*

$$\sum{}^{+}(X'_{1j} - \hat{\rho}\hat{h}_2)\left[\mathscr{R}\left(\frac{\hat{h}_2 - \hat{\rho}X'_{1j}}{\sqrt{1-\hat{\rho}^2}}\right)\right]^{-1} = \tag{214}$$

$$= \sum{}^{-}(X'_{1j} - \hat{\rho}\hat{h}_2)\left[\mathscr{R}\left(-\frac{h_2 - \hat{\rho}X'_{1j}}{\sqrt{1-\hat{\rho}^2}}\right)\right]^{-1},$$

$$\sum{}^{+}\left[\mathscr{R}\left(\frac{\hat{h}_2 - \hat{\rho}X'_{1j}}{\sqrt{1-\hat{\rho}^2}}\right]^{-1} = \sum{}^{-}\left[\mathscr{R}\left(\frac{\hat{h}_2 - \hat{\rho}X'_{1j}}{-\sqrt{1-\hat{\rho}^2}}\right)\right]^{-1}, \tag{215}$$

unde: $\sum^{+}$ *reprezintă suma după* j *pentru* $Y_j = 1$,
$\sum^{-}$ *reprezintă suma după* j *pentru* $Y_j = 0$;

$$X'_{1j} = \frac{X_{1j} - \bar{X}_1}{S} \quad \textit{și}$$

$$\mathscr{R}(u) = \frac{1 - \Phi(u)}{\varphi(u)} \textit{ (raportul lui Mills)}$$

Tate (1955) a construit o metodă iterativă de rezolvare a ecuațiilor (214) și (215) pentru $\hat{\rho}$ și $\hat{h}_2$. Sugerează să se ia ca valori inițiale ρ^* și h^*. Pentru valori mari ale lui n,

$$nD^2(\hat{h}_2) \simeq \frac{(1-\rho^2)(\psi_2 - 2\rho h_2\psi_1 + \rho^2 h_2\psi_0)}{\psi_0\psi_2 - \psi_1^2} + \rho^2(\rho^2 h_2^2 + 2), \tag{216}$$

$$nD^2(\hat{\rho}) \simeq \frac{(1-\rho^2)^3\psi_0}{\psi_0\psi_2 - \psi_1^2} + \rho^2(1-\rho^2)^2, \tag{217}$$

unde

$$\psi_r = \int_{-\infty}^{\infty} t^r\varphi(t)\left[\mathscr{R}\left(\frac{h_2 - \rho t}{\sqrt{1-\rho^2}}\right)\mathscr{R}\left(-\frac{h_2 - \rho t}{\sqrt{1-\rho^2}}\right)\right]^{-1} dt. \tag{218}$$

Scanned with OKEN Scanner

Prince, Tate (1966) dă valorile membrilor din dreapta ai relaţiilor (216) şi (217) şi funcţiilor ψ_0, ψ_1 şi ψ_2 cu 5 zecimale pentru $h_2 = 0(0{,}1)0{,}8(0{,}05)1{,}60(0{,}025)1{,}65$ şi diferite valori ale lui ρ.

Observaţie. (216) şi (217) nu se schimbă prin inversarea semnului lui ρ sau h, ψ_r nu se schimbă prin inversarea semnului lui ρ, însă se multiplică prin $(-1)^r$ dacă se schimbă semnul lui h_2.

§ 12. Repartiţii normale bidimensionale şi tridimensionale trunchiate

În acest paragraf vom folosi repartiţiile standardizate de densităţi (1) cu $\mu_1 = \mu_2 = 0$ şi $\sigma_1 = \sigma_2 = 1$ sau (161). Considerînd noua variabilă $\mu_t + \sigma_t X_t$, se obţine cazul general.

Cea mai folosită formă de trunchiere a repartiţiei normale bidimensionale este trunchierea simplă în raport cu una dintre variabile. Dacă alegem numai valorile variabilei X_1 care depăşesc h, repartiţia rezultată are densitatea de repartiţie :

$$p_{X_1,X_2}(x_1, x_2) = [1 - \Phi(h)]^{-1} \times \tag{219}$$

$$\times \exp\left[-\frac{1}{2(1-\rho^2)}(x_1^2 - 2\rho\, x_1 x_2 + x_2^2)\right], \quad x_1 > h.$$

Propoziţia 51. *Avem* :

$$M(X_2) = M[M[X_2|X_1]] = \rho M(X_1), \tag{220}$$

$$M[X_1X_2] = M[X_1 M[X_2|X_1]] = \rho M(X_1^2), \tag{221}$$

$$M(X_2^2) = M[M[X_2^2|X_1]] = M[\rho^2 X_1^2 + 1 - \rho^2] = \rho^2 M(X_1^2) + 1 - \rho^2. \tag{222}$$

Urmează că

$$D^2(X_2) = \rho^2 D^2(X_1) + 1 - \rho^2 \tag{223}$$

$$\operatorname{cov}(X_1, X_2) = \rho D^2(X_1), \tag{224}$$

şi

$$\rho(X_1, X_2) = \rho\sqrt{D^2(X_1)/D^2(X_2)} = \rho\left[\rho^2 + \frac{1-\rho^2}{D^2(X_1)}\right]^{-1/2}. \tag{225}$$

(225) este adevărată pentru orice formă de trunchiere a variabilei X_1 cu condiţia ca X_2 să nu fie trunchiată.

Avem :

$$D^2(X_1) = 1 + \frac{h\varphi(h)}{1-\Phi(h)} - \left\{\frac{\varphi(h)}{1-\Phi(h)}\right\}^2. \tag{226}$$

Deoarecere

$$D^2(X_1) \leqslant 1$$

urmează că

$$|\rho(X_1, X_2)| \leqslant |\rho|. \tag{227}$$

Astfel coeficientul de corelație în populația trunchiată este numeric mai mic în populația originală. Tabelul 5 dă cîteva valori numerice.

Tabelul 5

Coeficientul de corelație în populația normală trunchiată

Gradul de trunchiere $\Phi(h)$	Coef. de corelație din populația netrunchiată ρ			
	0,25	0,5	0,75	0,9
0,1	0,213	0,438	0,691	0,867
0,2	0,193	0,403	0,655	0,845
0,3	0,178	0,376	0,623	0,823
0,4	0,165	0,351	0,593	0,802
0,5	0,154	0,329	0,564	0,780
0,6	0,143	0,307	0,535	0,755
0,7	0,132	0,285	0,504	0,728
0,8	0,120	0,261	0,468	0,695
0,9	0,106	0,231	0,423	0,647

PROPOZIȚIA 52. *În populația normală bidimensională trunchiată, avem*

$$M[X_1|X_2] = \rho X_2 + \frac{\varphi[(h - \rho X_2)/(\sqrt{1-\rho^2})]}{1 - \Phi[(h - \rho X_2)/\sqrt{1-\rho^2}]} \cdot \sqrt{1-\rho^2} \tag{228}$$

PROPOZIȚIA 53. *Dacă X_1 și X_2 sînt trunchiate la dreapta ($X_1 > h$, $X_2 > h$), atunci momentele (μ'_{rs}) sînt date de*

$$L(h, k; \rho)\mu'_{10} = \varphi(h)\{1 - \Phi(A)\} + \rho\varphi(k)\{1 - \Phi(B)\}, \tag{229}$$

$$L(h, k; \rho)\mu'_{20} = h\varphi(h)\{1 - \Phi(A)\} + \rho^2 k\varphi(k)\{1 - \Phi(B)\} + \tag{230}$$

$$+ \rho(1 - \rho^2)\frac{1}{2}\varphi(h, k; \rho) + L(h, k; \rho)$$

$$L(h, k; \rho)\mu'_{11} = \rho[h\varphi(h)\{1 - \Phi(A)\} + k\varphi(k)\{1 - \Phi(B)\} + \tag{231}$$

$$+ L(h, k; \rho) + (1 - \rho^2)\varphi(h, k; \rho),$$

unde

$$A = \frac{k - \rho h^2}{\sqrt{1-\rho^2}}; \quad B = \frac{h - \rho k^2}{\sqrt{1-\rho^2}} \tag{232}$$

$$\varphi(h, k; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}(h^2 - 2\rho hk + k^2)\right]. \tag{233}$$

Scanned with OKEN Scanner

Observație. Formulele pentru μ'_{or} sînt obținute din acelea pentru μ'_{ro} schimbînd h cu k, A și B.

PROPOZIȚIA 54 (Rosenbaum (1961)). *Avem*

$$(h+k)\rho^2 - \{(h+k)\mu'_{11} - hk(\mu'_{10}+\mu'_{01})\}\rho -$$

$$-(h+k) - hk(\mu'_{10}+\mu'_{01}) + k\mu'_{20} + h\mu'_{02} = 0. \tag{234}$$

Rezultă din (229) — (231).

Cazul în care amîndouă variabilele sînt trunchiate a fost considerat de Des Raj (1953), Shah, Parikh (1964) și Regier, Hamdan (1971).

PROPOZIȚIA 55 (Shah, Parikh (1964)). *Dacă*

$$h_1 < X_1 < k_1,\ h_2 < X_2 < k_2$$

atunci (*cu* $\mu'_{r,s} = M[X_1^r X_2^s]$)

$$\mu'_{r,s} - (r-1)(1-\rho^2)\mu'_{r-2,s} - \rho\mu'_{r-1,s+1} =$$

$$= P^{-1}(1-\rho^2)[h_1^{r-1}\varphi(h_1)G_s(h_2, k_2, h_1\rho;\ \sqrt{1-\rho^2})] - \tag{235}$$

$$- k_1^{r-1}\varphi(k_1)G_s(h_2, k_2, k_1\rho;\ \sqrt{1-\rho^2})],\ (r \geqslant 2s,\ s \geqslant 2)$$

unde

$$P = \int_{h_1}^{h_2}\int_{k_1}^{k_2} \varphi(0,0;\ 1,1;\ \rho)\mathrm{d}x_1\mathrm{d}x_2 \tag{236}$$

și

$$G_s(a_1, a_2, b;\ c) = c^{-1}\int_{a_1}^{a_2} x^s \varphi\left(\frac{x-b}{c}\right)\mathrm{d}x. \tag{237}$$

De asemenea:

$$\mu'_{r-1,s+1} - s(1-\rho^2)\mu'_{r-1,s-1} - \rho\mu'_{r,s} =$$

$$= P^{-1}(1-\rho^2)[h_2^s\varphi(h_2)G_{r-1}(h_1, k_1, h_2\rho;\ \sqrt{1-\rho^2}) - \tag{238}$$

$$- k_2^s\varphi(k_2)G_{r-1}(h_1, k_1, k_2\rho;\ \sqrt{1-\rho^2})].$$

În cazul în care amîndouă variabilele sînt trunchiate în același punct (astfel încît sînt reținute valorile $X_1 > a$, $X_2 > a$), Regier, Hamdan (1971) dă valori ale coeficientului de corelație ρ' ca funcție de a și de coeficientul de corelație ρ dinaintea trunchierii cu 3 zecimale pentru $\pm a = 0{,}0(0{,}1)1{,}1(0{,}2)1{,}5(0{,}5)2{,}5$; $\rho = 0{,}05(0{,}05)0{,}95$.

Lipow, Eidemiller (1964) [Parikh, Sheth)1966)] evaluează $P(X_1 - X_2 > 0)$ pentru trunchierea simplă (dublă) folosind tabelele cu funcția $L(h, k;\ \rho)$.

În cazul repartiției normale tridimensionale are loc.

Scanned with OKEN Scanner

PROPOZIȚIA 56. *Dacă este trunchiată o singură variabilă* $(X_1 \geqslant h)$, *atunci*:

$$M(X_t) = \rho_{1t} M(X_1); \quad M(X_t^2) = \rho_{1t}^2 M(X_t^2) + 1 - \rho_{1t}^2;$$

$$D^2(X_t) = \rho_{1t}^2 D^2(X_1) + 1 - \rho_{1t}^2; \tag{239}$$

$$\rho(X_1, X_t) = \rho_{1t}\left[\rho_{1t}^2 + \frac{1-\rho_{1t}^2}{D^2(X_1)}\right]^{-1/2}; \quad t = 2, 3 \tag{240}$$

$$M(X_2X_3) = M[M[X_2X_3|X_1]] = M[\rho_{231}\sqrt{(1-\rho_{12}^2)(1-\rho_{13}^2)} + \rho_{12}\rho_{13}X_1^2] = \tag{241}$$

$$= \rho_{12}\rho_{13}M(X_1^2) + \rho_{23} - \rho_{12}\rho_{13}.$$

Rezultă din Propoziția 43.
Urmează că

$$\operatorname{cov}(X_2, X_3) = \rho_{12}\rho_{13}D^2(X_1) + \rho_{23} - \rho_{12}\rho_{13} \tag{242}$$

și

$$\rho(X_2, X_3) = \frac{\rho_{23} - \rho_{12}\rho_{13}\left(\frac{1}{D^2(X_1)} - 1\right)}{\sqrt{\left(\rho_{12}^2 + \frac{1-\rho_{12}^2}{D^2(X_1)}\right)\left(\rho_{13}^2 + \frac{1-\rho_{13}^2}{D^2(X_1)}\right)}}. \tag{243}$$

Dacă trunchierea se face alegînd numai valorile X_1, X_2, astfel încît $\alpha_1X_1 + \alpha_2X_2 > h$, problema se reduce la aceea a repartiției normale bidimensionale cu trunchierea unei singure variabile. Aceasta se întîmplă deoarece $(\alpha_1X_1 + \alpha_2X_2)$ și X_3 înainte de trunchiere) au o repartiție normală bidimensională. Dacă mai avem o trunchiere, $\alpha_1'X_1 + \alpha_1'X_2 > h'$, atunci printr-o transformare, obținem trunchieri de forma $Z > h$, $Z' > h'$ Birnbaum (1950), Young, Weiler (1960))

§ 13. Corelații canonice

13.1. Populații normale

DEFINIȚIA 17. Dacă mulțimea variabilelor $\mathbf{X}' = (X_1, \ldots, X_m)$ este împărțită în două submulțimi $\mathbf{X}_{(1)}' = (X_1, \ldots, X_{m_1})$ și $\mathbf{X}_{(2)}' = (X_{m_1+1}, \ldots, X_m)$, atunci rădăcinile ecuației

$$|\boldsymbol{\Sigma}_{11} - \rho^2\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}| = 0 \tag{224}$$

unde

$$\boldsymbol{\Sigma} = \begin{vmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{vmatrix}$$

cu $\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{22}$ matricile de covarianțe ale vectorilor $\mathbf{X}'$, $\mathbf{X}_{(1)}'$, respectiv $\mathbf{X}_{(2)}'$ poartă numele de corelațiile canonice dintre $\mathbf{X}_{(1)}'$ și $\mathbf{X}_{(2)}'$.

Scanned with OKEN Scanner

Întotdeauna se poate alege ca $m_1 \leqslant m - m_1 = m_2$ astfel că în general există m_1 valori pozitive $\rho_1^2 \leqslant \rho_2^2 \leqslant \ldots \leqslant \rho_{m_1}^2$ ale lui ρ satisfăcînd (244).

TEOREMA 15. (Constantine (1963)). *Valorile* $R_1^2 \leqslant R_2^2 \leqslant \ldots \leqslant R_{m_1}^2$ *ce satisfac ecuaţia*

$$|\mathbf{S}_{11} - r^2\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}| = \mathbf{0} \tag{245}$$

unde

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix}$$

are o repartiţie Wishart cu ν *grade de libertate şi matricea de covariante* $\mathbf{\Sigma}$, *au repartiţia comună*

$$p_{R^2}(r^2) = \frac{\pi^{(1/2)m_1^2}\,\Gamma_{m_1}\left(\frac{1}{2}\nu\right)}{\Gamma_{m_1}\left(\frac{1}{2}m_1\right)\Gamma_{m_1}\left(\frac{1}{2}m_2\right)\Gamma_{m_1}\left(\frac{1}{2}(\nu - m_2)\right)} \times$$

$$\times \prod_{j=1}^{m_1} [(1 - \rho_j^2)^{(1/2)\nu}(r_j^2)^{(1/2)(m_2 - m_1 - 1)}(1 - r_j^2)^{(1/2)(\nu - m - 1)}] \times \tag{246}$$

$$\times \left[\prod_{i>j}^{m_1}(r_i^2 - r_j^2)\right] {}_2F_1\left(\frac{1}{2}\nu; \frac{1}{2}\nu; \frac{1}{2}m_2; \mathbf{P}^2, \mathbf{D}^2\right),$$

$$(0 \leqslant r_1^2 \leqslant \ldots \leqslant r_{m_1}^2 \leqslant 1),$$

unde $\mathbf{P}^2 = \operatorname{diag}(\rho_1^2, \ldots, \rho_{m_1}^2)$; $\mathbf{D}^2 = \operatorname{diag}(r_1^2, \ldots, r_{m_1}^2)$.

Observaţie. Dacă $\mathbf{P} = \mathbf{0}$ (independenţă completă în cazul normal) (246) se reduce la

$$p_{R^2}(r^2) = \frac{\pi^{(1/2)m_1^2}\,\Gamma_{m_1}\left(\frac{1}{2}\nu\right)}{\Gamma_{m_1}\left(\frac{1}{2}m_1\right)\Gamma_{m_1}\left(\frac{1}{2}m_2\right)\Gamma_{m_1}\left(\frac{1}{2}(\nu - m_2)\right)} \times$$

$$\times \prod_{j=1}^{m_1} [(r_j^2)^{(1/2)(m_2 - m_1 - 1)}(1 - r_j^2)^{(1/2)(\nu - m - 1)}] \prod_{i>j}^{m_1}(r_i^2 - r_j^2), \tag{247}$$

$$(0 \leqslant r_1^2 \leqslant \ldots \leqslant r_{m_1}^2 \leqslant 1).$$

Pillai, Sugiyama (1969) şi de Waal (1969) au dat expresii (complicate) pentru funcţia de repartiţie a pătratului celei mai mari corelaţii canonice.

Scanned with OKEN Scanner

TEOREMA 15. (De Waal (1969)). *Avem*:

$$P[R^2_{m_1} \leqslant \tau^2] = \frac{\Gamma_{m_1}\left(\frac{1}{2}\nu\right)\Gamma_{m_1}\left(\frac{1}{2}(m_1+1)\right)}{\Gamma_{m_1}\left(\frac{1}{2}(\nu-m)\right)\Gamma_{m_1}\left(\frac{1}{2}(m_1+m+1)\right)}\left(\prod_{j=1}^{m_1}(1-\rho_j^2)\right)^{(1/2)\nu} \times$$

$$\times \sum_k \sum_K \sum_j \sum_J \sum_\delta (\tau^2)^{(1/2)mm_1+k+j} A^\delta_{K,J}(p_1^2, \ldots, p_{m_1}^2),$$

unde j şi k sînt sumaţi de la 0 la ∞, $A^\delta_{K,J}$ constante cu expresii greu de calculat, J, K sînt partiţii ale j şi k şi δ de $j+k$, cu cel puţin m_1+1 componente.

13.2. Populaţii care nu sînt normale

Muirhead, Waternaux (1980) notează rădăcinile caracteristice ale matricei $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}'$ prin

$$\rho_1^2, \ldots, \rho_{m_1}^2 (1 > \rho_1^2 \geqslant \ldots \geqslant \rho_{m_1}^2 \geqslant 0) \tag{249}$$

şi ale matricei $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{12}'$ prin

$$r_1^2, \ldots, r_{m_1}^2 (1 > r_1^2 \geqslant \ldots \geqslant r_p^2 > 0). \tag{250}$$

Putem scrie

$$r_i^2 = a_{ii} + \sum_{j \neq i} \frac{a_{ij}a_{ji}}{\rho_i^2 - \rho_j^2} + \mathbf{0}_p(N^{-3/2}), \tag{251}$$

unde $\mathbf{A} = ((a_{ij})) = \mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{12}'$ (Lawley (1959), Muirhead, Waternaux (1980)) şi N este volumul selecţiei.

Cumulanţii de ordinul patru ai populaţiei vor fi notaţi prin

$$\mathfrak{K}_{1111}^{ijkl} = M(x^{(i)}x^{(j)}x^{(k)}x^{(l)}) - \sigma_{ij}\sigma_{kl} - \sigma_{ik}\sigma_{jl} - \sigma_{il}\sigma_{jk}. \tag{252}$$

Excesul componentei i este

$$\mathfrak{K}_i^i = M(x^{(i)4}) - 3 \tag{253}$$

şi cumulanţii de ordinul patru bidimensionali sînt

$$\begin{aligned} \mathfrak{K}_{22}^{ij} &= \operatorname{cov}(x^{(i)2}, x^{(j)2}) = M(x^{(i)2}x^{(j)2}) - 1 \quad (j \neq i + m_1), \\ \mathfrak{K}_{22}^{i,i+m_1} &= M(x^{(i)2}x^{(i+m_1)2}) - 1 - 2\rho_i^2. \end{aligned} \tag{254}$$

Scanned with OKEN Scanner

Fără a pierde din generalitate putem presupune că $M(X) = 0$ și că matricea de covarianță $\mathbf{\Sigma}$ are forma

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{I}_{m_1} & \mathbf{P} \\ \mathbf{P} & \mathbf{I}_{m_2} \end{pmatrix} \tag{255}$$

unde $\mathbf{P} = [\operatorname{diag}(\rho_1, \ldots, \rho_{m_1}) : 0]$ este $m_1 \times m_2$

Cu

$$\begin{cases} \mathbf{Z}_{11} = N^{1/2}(\mathbf{I}_{m_1} - \mathbf{PP}')^{-1/2}(\mathbf{S}_{11} - \mathbf{I}_{m_1})(\mathbf{I}_{m_1} - \mathbf{PP}')^{-1/2}, \\ \mathbf{Z}_{22} = N^{1/2}(\mathbf{I}_{m_2} - \mathbf{P}'\mathbf{P})^{-1/2}(\mathbf{S}_{22} - \mathbf{I}_{m_2})(\mathbf{I}_{m_2} - \mathbf{P}'\mathbf{P})^{-1/2}, \\ \mathbf{Z}_{12} = N^{1/2}(\mathbf{I}_{m_1} - \mathbf{PP}')^{-1/2}(\mathbf{S}_{12} - \mathbf{P})(\mathbf{I}_{m_2} - \mathbf{P}'\mathbf{P})^{1/2}, \end{cases} \tag{256}$$

$$\mathbf{Z} = ((z_{ij})) = \begin{pmatrix} \mathbf{Z}_{11} & \mathbf{Z}_{12} \\ \mathbf{Z}'_{12} & \mathbf{Z}_{22} \end{pmatrix},$$

$$A_i = \rho_i^2(1 - \rho_i^2)(2z_{i,m_1+i} - \rho_i z_{ii} - \rho_i z_{m_1+i,m_1+i}), \tag{257}$$

$$\begin{aligned} B_i = {} & \alpha_i \rho_i^2 z_{ii}(z_{ii} + z_{m_1+i,m_1+i}) - 2\alpha_i \rho_i z_{i,m_1+i}(z_{ii} + z_{i,m_1+1}) + \\ & + \alpha_i(z_{i,m_1+1}^2 + \rho_i^2 z_{m_1+i,m_1+i}^2) + \\ & + \sum_{j \neq i+1}^{m_1} \beta_{ij}\{\rho_i^3(z_{ij}^2 + z_{m_1+i,m_1+j}^2) + \rho_i(z_{i,m_1+j}^2 + z_{jm_1+i}^2) - \\ & - 2\rho_i\rho_j(z_{ij}z_{i,m_1+j} + z_{j,m_1+i}z_{m_1+i,m_1+j}) - \\ & - 2\rho_i^2(z_{ij}z_{j,m_1+i} + z_{i,m_1+j}z_{m_1+i,m_1+j}) + 2\rho_j(\rho_j^2 z_{ij}z_{m_1+i,m_1+j} + \\ & + z_{i,m_1+j}z_{j,m_1+i})\} + \gamma_i \sum_{j=p+1}^{m_2} (z_{i,m_1+j}^2 + \rho_i^2 z_{m_1+i,m_1+j}^2 - 2\rho_i z_{i,m_1+j}z_{m_1+i,m_1+j}), \end{aligned} \tag{258}$$

$$\gamma_i = 1 - \rho_i^2,\ \alpha_i = (1 - \rho_i^2)^2,\ \beta_{ij} = \rho_i(1 - \rho_i^2)(1 - \rho_j^2)/(\rho_i^2 - \rho_j^2), \tag{259}$$

avem

$$r_i^2 = \rho_i^2 + N^{-1/2}A_i + N^{-1}B_i + \mathbf{0}_p(N^{-3/2}). \tag{260}$$

Are loc

TEOREMA 16. (Muirhead, Waternaux (1980)). *Fie r_i $(i = 1, \ldots, m_1)$ coeficienții de corelație canonică a unei selecții de volum N dintr-o populație având cumulanții de ordinul patru $\varkappa_{1111}^{ijkl}$ finiți. Dacă coeficienții de corelație canonică ai populației sînt distincți, repartiția comună limită, cînd $N \to \infty$, a statisticii*

$$\sqrt{N}(r_i^2 - \rho_i^2)(i = 1, \ldots, m_1)$$

este normală p-dimensională cu vectorul valoare medie $\mathbf{0}$ și matricea de cova-

riante $\Lambda = ((\lambda_{ij}))$ *avînd elementele*:

$$\lambda_{ii} = 4\rho_i^2(1-\rho_i^2)^2 + \rho_i^4(\mathcal{K}_4^i + \mathcal{K}_4^{m_1+i}) + 2\rho_i^2(\rho_i^2+2)\mathcal{K}_{22}^{i,m_1+i} -$$

$$- 4\rho_i^3(\mathcal{K}_{31}^{i,m_1+1} + \mathcal{K}_{13}^{i,m_1+1}), \quad (i = 1, \ldots, p),$$

$$\lambda_{ij} = \rho_i^2\rho_j^2(\mathcal{K}_{22}^{ij} + \mathcal{K}_{22}^{i,m_1+j} + \mathcal{K}_{22}^{j,m_1+i} + \mathcal{K}_{22}^{m_1+i,m_1+j}) -$$

$$- 2\rho_i^2\rho_j(\mathcal{K}_{211}^{ij,m_1+j} + \mathcal{K}_{211}^{m_1+i,m_1+j}) - 2\rho_i\rho_j^2(\mathcal{K}_{211}^{j,i,m_1+i} + \mathcal{K}_{211}^{m_1+j,ij}) +$$

$$+ 4\rho_i\rho_j\mathcal{K}_{1111}^{ij,m_1+i,m_1+j} \ (i, j = 1, \ldots, m_1; \ i \neq j).$$

DEFINIŢIA 18. Vom spune că $\mathbf{X}$ are o repartiţie eliptică dacă are densitatea de repartiţie

$$c\,|\Sigma|^{-1/2} g(\mathbf{x}'\Sigma^{-1}\mathbf{x}), \tag{261}$$

unde Σ este o matrice simetrică pozitiv definită şi g o funcţie nenegativă.

Dacă există matricea de covarianţe, ea este proporţională cu Σ; dacă $\Sigma = \lambda\mathbf{I}$, repartiţia este numită sferică. Repartiţia normală n-dimensională cu densitatea

$$\phi(\mathbf{x}\,|\Sigma) = (2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}\mathbf{x}'\Sigma^{-1}\mathbf{x}\right) \tag{262}$$

şi repartiţia normală ε-contaminată cu densitatea

$$\phi_\varepsilon(\mathbf{x}\,|\,\sigma, \Sigma) = (1-\varepsilon)\phi(\mathbf{x}\,|\Sigma) + \frac{\varepsilon}{\sigma^n}\phi(\mathbf{x}/\sigma\,|\Sigma) \tag{263}$$

sînt membrii ai acestei clase.

Pentru repartiţiile eliptice toţi cumulanţii de ordinul patru sînt determinaţi de un parametru $\mathcal{K}$ care caracterizează excesul repartiţiei şi care depinde numai de g. Se defineşte $\mathcal{K}$ prin $\mathcal{K}_4^i = 3\mathcal{K}$ $(i = 1, \ldots, m_1 + m_2)$ unde $\mathcal{K}_4^i$ este excesul repartiţiei marginale a lui x_i, sau excesul standardizat dacă $D^2(x_i) \neq 1$. Pentru repartiţia normală ε-contaminată cu densitatea $\phi_\varepsilon(\mathbf{x}\,|\,\sigma, \Sigma)$

$$\mathcal{K} = \frac{1 + \varepsilon(\sigma^4 - 1)}{\{1 + \varepsilon(\sigma^2 - 1)\}^2} - 1. \tag{264}$$

Orice repartiţie eliptică cu excesul standardizat $3\mathcal{K}$ are cumulanţii de ordinul patru

$$\mathcal{K}_{1111}^{ijkl} = \mathcal{K}(\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}). \tag{265}$$

Pentru cazul particular (255),

$$\mathcal{K}_{22}^{ij} = \mathcal{K}(j \neq i, i + m_1), \ \mathcal{K}_{22}^{i,i+m_1} = (1 + 2\rho_i^2)\mathcal{K},$$

$$\mathcal{K}_{31}^{i,i+m_1} = \mathcal{K}_{13}^{i,i+m_1} = 3\rho_i\mathcal{K}, \ \mathcal{K}_{1111}^{i,i+m_1,j,j+m_1} = \rho_i\rho_j\mathcal{K},$$

şi toţi cumulanţii de ordinul patru sînt zero,

Scanned with OKEN Scanner

COROLARUL 1. (Muirhead, Waternaux (1980)). *Fie* $r_i (i=1, \ldots, p)$ *coeficienţii de corelaţie canonică a unei selecţii de volum* N *asupra unei variabile eliptice cu excesul* $3\mathcal{K}$. *Fie* $\mathbf{u}$ *vectorul* $m_1 \times 1$ *avînd componentele*

$$u_i = \sqrt{N}\,\frac{r_i^2 - \rho_i^2}{2\rho_i(1-\rho_i^2)} (i = 1, \ldots, m_1).$$

Dacă coeficienţii de corelaţie canonică ai populaţiei sînt distincţi, repartiţia limită a lui $\mathbf{u}$, *cînd* $N \to \infty$, *este normală multidimensională cu vectorul valoare medie* $\mathbf{0}$ *şi matricea de covarianţe* $(\mathcal{K}+1)\mathbf{I}_p$.

Observaţie. Cînd observaţiile provin dintr-o populaţie normală multidimensională ($\mathcal{K} = 0$) u_i sînt asimptotic independente şi repartizate $N(0,1)$ [Hsu (1941)].

§ 14. Coeficientul de corelaţie pentru date direcţionale

O problemă naturală ce se ridică în statistica direcţională este aceea a definiţiei coeficientului de corelaţie dintre două variabile ce iau valori în subspaţii generale. De o importanţă particulară este cazul bidirecţional în care ambele variabile variază pe un cerc. Au fost definiţi coeficienţii de corelaţie circulari, mulţi dintre aceştia fiind generalizaţi pentru alte subspaţii.

Fie $\mathbf{X}$ şi $\mathbf{Y}$ doi vectori aleatori cu matricea de covarianţe $\mathbf{\Sigma}$ partiţionată ca :

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}$$

şi presupunem că $\mathbf{\Sigma}_{11}$ *şi* $\mathbf{\Sigma}_{22}$ *sînt nesingulare de ranguri* p *şi* q $(p \leqslant q)$.

Pentru cerc şi sferă are loc

DEFINIŢIA 19 (Downs (1974), Mardia 1975)). Coeficientul de corelaţie ρ_1^2 dintre $\mathbf{X}$ şi $\mathbf{Y}$ este

$$\rho_1^2 = \frac{\{Ur(\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{21})^{1/2}\}^2}{Ur(\mathbf{\Sigma}_{11})\,Ur(\mathbf{\Sigma}_{22})}. \tag{266}$$

PROPOZIŢIA 57. i) ρ_1^2 *este invariant*

ii) $\rho_1^2 \leqslant 1$ *cu egal dacă şi numai dacă*

$$\mathbf{t}_1 - M(\mathbf{t}_1) = \lambda U\{\mathbf{t}_2 - M(\mathbf{t}_2)\},$$

unde λ *este o constantă şi* $\mathbf{UU}' = \mathbf{I}_p$

Downs, Eifler (1975) arată că repartiţia nulă a statistlcii nr_1^2 depinde de repartiţiile marginale.

Cu $\mathbf{\Sigma}_{ij}^{+}$ inversa generalizată Penrose a lui $\mathbf{\Sigma}_{ij}$, pentru cerc şi sferă are loc :

DEFINIŢIA 20 (Mardia, Puri (1978)). Coeficientul de corelaţie ρ_2^2 este :

$$\rho_2^2 = Ur(\hat{\mathbf{\Sigma}}_{11}^{+}\mathbf{\Sigma}_{12}\hat{\mathbf{\Sigma}}_{22}^{+}\mathbf{\Sigma}_{21}), \tag{267}$$

Scanned with OKEN Scanner

unde $\hat{\Sigma}_{ii}(i = 1, 2)$ se referă la expresiile pentru Σ_{ii} obținute din $M(\mathbf{t}_i)$ prin metoda V.M. în familia exponențială cu densitatea

$$\exp\{\mathbf{A}'\mathbf{t}_i - \mathcal{K}(\mathbf{A})\},$$

(familia von Mises sau Fisher în cazul circular sau sferic).

PROPOZIȚIA 58. *Avem*:

$$\rho_2^2 = \sum_{i,j=1}^{2} \{\mathrm{cov}\,(t_{1i}, t_{2j})\}^2/(a_{1i}a_{2j}),$$

unde, pentru $k = 1, 2$,

$$\mathbf{t}_1'(\theta) = (\cos\theta, \sin\theta),\ \mathbf{t}_2'(\theta) = (\cos\phi, \sin\phi),$$

$$a_{k1} = 1 - a_{k2} - \{M(t_{k1})\}^2,\ a_{k2} = M(t_{k1})/A^{-1}\{M(t_{k1})\},$$

$$\mathbf{A}(\mathcal{K}) = \mathbf{I}_1(\mathcal{K})/\mathbf{I}_0(\mathcal{K}).$$

PROPOZIȚIA 59. (Mardia, Puri (1978)). *Asimptotic, în cazul repartițiilor marginale* von Mises, $nr_2^2 \sim \chi^2(pq)$.

DEFINIȚIA 21. (Mardia, Puri (1978)). Coeficientul de corelație ρ_3^2 este

$$\rho_3^2 = Ur(\check{\Sigma}_{11}^{+}\Sigma_{12}\check{\Sigma}_{22}^{+}\Sigma_{21}), \tag{268}$$

unde $\mathbf{V}\check{\Sigma}_{11}\mathbf{V}'$ este diagonală cu aceleași elemente diagonale ca și $\mathbf{V}\Sigma_{11}\mathbf{V}'$ pentru o matrice ortogonală $\mathbf{V}$ convenabil aleasă.

Dacă rang $\Sigma_{11} = 1$ orice alegere a lui $\mathbf{V}$ este convenabilă.

Dacă rang $\Sigma_{11} = 2$ și μ_0 media direcției lui $\mathbf{t}_1$ există, atunci μ_0 este prima linie a lui $\mathbf{V}$.

Dacă $\mathbf{t}_1 = \mathbf{t}_2$, rang Σ_{11} = rang $\Sigma_{22} = 3$ și dacă μ_0 și ν_0 (media direcției lui $\mathbf{t}_2$) există cu $\mu_0'\nu_0 \neq \pm 1$, atunci μ_0 este prima linie și $\nu_0 - (\mu_0'\nu_0)\mu_0$ este multiplu de a doua linie a lui $\mathbf{V}$.

Matricea $\check{\Sigma}_{22}$ este similar definită.

Nu este clar cum alegem $\check{\Sigma}_{11}$ și $\check{\Sigma}_{22}$ în alte cazuri.

PROPOZIȚIA 60. $\rho_3^2 \leqslant p_1p_2$, *unde* p_i *este dimensiunea lui* $\mathbf{t}_i$ *pentru* $i = 1, 2$, *cu egal dacă numai dacă*

$$\text{rang } \Sigma_{11} = \text{rang } \Sigma_{22} = 1$$

și

$$\mathbf{t}_2(Y) = a\mathbf{t}_1(X) + b$$

a și b constante.

În cazul circular bidimensional are loc

DEFINIȚIA 22 (Johnson, Wehrly (1977)). Coeficientul de corelație ρ_4^2 este definit ca cea mai mare valoare proprie a matricei $\Sigma_{11}^{+}\Sigma_{12}\Sigma_{22}^{+}\Sigma_{21}$.

PROPOZIȚIA 61. $\rho_4^2 \leqslant 1$.

Cu toate că în general $\rho_4^2 = 1$ nu implică dependența funcțională, în cazul bidimensional circular implică dependența rotațională de forma $\phi = \pm\theta - \alpha$.

Scanned with OKEN Scanner

Nu se cunoaşte dacă repartiţia asimptotică a statisticii nr_4^2 depinde de repartiţiile marginale.

DEFINIŢIA 23 (Thompson (1975)). Coeficientul de corelaţie este definit ca

$$\rho_5^2 = (\textit{coeficientul de corelaţie dintre } \theta^* \textit{ şi } \phi^*)^2 \tag{269}$$

unde

$$\theta^* = \theta - \mu_0 + \pi, \mod(2\pi), \quad \phi^* = \phi - \nu_0 + \pi, \mod(2\pi)$$

cu condiţia ca μ_0 şi ν_0 să existe.

De fapt este vorba de o modificare a coeficientului lui Thompson obţinută înlocuind $\sqrt{2}\sin\left(\frac{1}{2}\theta^*\right)$ prin θ^*.

Pentru repartiţii absolut continui, r_5^2 este definit cu probabilitatea unu.

Nu este cunoscut dacă repartiţia asimptotică (nulă) a statisticii nr_5^2 depinde de repartiţiile marginale.

În cazul sferic are loc

DEFINIŢIA 24. (Mackenzie (1957), Stephens (1979)). *Coeficientul de corelaţie* ρ_6^2 *este definit ca*

$$\rho_6^2 = Ur\{(\mathbf{AA}')^{1/2}\}, \quad \mathbf{A} = M(\mathbf{t}_1\mathbf{t}_2'). \tag{270}$$

PROPOZIŢIA 62 (Stephens (1979)). ρ_6^2 *este invariant la transformări ortogonale; independenţa lui* $\mathbf{X}$ *şi* $\mathbf{Y}$ *nu implică că* $\rho_6^2 = 0$.

Repartiţia asimptotică (nulă) a statisticii r_6^2 depinde de repartiţiile marginale.

DEFINIŢIA 25 (Jupp, Mardia (1980)). Coeficientul de corelaţie ρ^2 dintre $\mathbf{X}$ şi $\mathbf{Y}$ este definit ca:

$$\rho^2 = Ur(\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}). \tag{271}$$

PROPOZIŢIA 63. *Avem*:

$$\rho^2 = \sum_{i=1}^{s} \lambda_i^2, \tag{272}$$

unde $\lambda_1, \ldots, \lambda_s$ *sînt corelaţiile canonice şi* $s = min\,(p, q)$.

În loc de a defini ρ^2 prin (272) se poate folosi statistica Hotelling-Lawley

$$\sum_{i=1}^{s} \lambda_i^2/(1 - \lambda_i^2)$$

sau statistica Wilks

$$\prod_{i=1}^{s} (1 - \lambda_i^2).$$

Privind cercul ca inclus în plan putem folosi (271) pentru a defini coeficientul de corelaţie pentru cazul circular bidimensional.

Scanned with OKEN Scanner

Dacă θ și ϕ sînt variabilele circulare, $0 \leqslant \theta$, $\phi \leqslant 2\pi$ și punem :

$$t_1'(\theta) = (\cos\theta, \sin\theta), \quad t_2'(\theta) = (\cos\phi, \sin\phi),$$

atunci definim ρ^2 prin (271), unde Σ este matricea de covarianțe a lui t_1 și t_2. Explicit avem :

$$\rho^2 = \{\rho_{cc}^2 + \rho_{cs}^2 + \rho_{sc}^2 + \rho_{ss}^2 + 2(\rho_{cc}\rho_{ss} - \rho_{cs}\rho_{sc})\,\rho_1\rho_2 -$$

(273)

$$-2(\rho_{cc}\rho_{cs} + \rho_{sc}\rho_{ss})\,\rho_1 - 2(\rho_{cc}\rho_{sc} + \rho_{cs}\rho_{ss})\,\rho_2\}/\{(1 - \rho_1^2)(1 - \rho_2^2)\},$$

unde ρ_{cc} = coef. de corel. dintre cos θ și cos ϕ, ...

ρ_1 = coef. de corel. dintre cos θ și sin θ, ...

Proprietatea de invarianță a coeficientului ρ^2 implică faptul că ρ^2 rămîne neschimbat alegînd,

$$t_1'(\theta) = \{(\cos(\theta + \alpha), \sin(\theta + \alpha)\}.$$

Similar pentru coeficienții de corelație sferici, privim sfera ca inclusă în R^3.

În cazul unei variabile aleatoare X pe subspațiul Stiefel $V_n(R^p) = O(n, p)$ al matricilor $\mathbf{X}$, $n \times p$ cu $\mathbf{XX}' = \mathbf{I}_n$(considerată de Downs (1972) și Khatri, Mardia (1977)) folosim $\mathbf{t}_1(\mathbf{X}) = \mathbf{X}$ în mulțimea tuturor matricilor $n \times p$, privită ca R^{np}.

În general, fie $\mathbf{X}$ și $\mathbf{Y}$ variabile aleatoare pe M și N, unde fiecare dintre M și N poate fi spațiu vectorial Euclidian sau subspațiul compact Riemannian.

Dacă M este spațiu vectorial punem $\mathbf{t}_1(\mathbf{X}) = \mathbf{X}$. Similar pentru $\mathbf{t}_2$. Fie Σ matricea de covarianțe a lui $\mathbf{t}_1$ și $\mathbf{t}_2$. Atunci are loc

DEFINIȚIA 26. Coeficientul de corelație este definit de

$$\rho^2 = Ur(\Sigma_{11}^{+}\Sigma_{12}\Sigma_{22}^{+}\Sigma_{21}). \tag{274}$$

Dacă $M = N$ și $\mathbf{t}_1 = \mathbf{t}_2$, putem alege ρ ca să aibă semnul lui $\det \Sigma_{12}$.

În cazul circular-liniar al unei variabile liniare x și a unei variabile unghiulare θ, are loc :

DEFINIȚIA 27 (Mardia (1976)). Coeficientul de corelație este definit de

$$\rho^2 = (\rho_{xc}^2 + \rho_{xs}^2 - 2\rho_{xc}\rho_{xs}\rho_{cs})/(1 - \rho_{cs}^2) \tag{275}$$

unde ρ_{xc} = coef. de corel. dintre x și cos θ

ρ_{cs} = coef. de corel. dintre cos θ și sin θ

. . . . . .

DEFINIȚIA 28 (Jupp, Mardia (1980)). Coeficientul de corelație de selecție r^2, bazat pe o selecție de volum n, este definit ca :

$$r^2 = Ur(\mathbf{S}_{11}^{+}\mathbf{S}_{12}\mathbf{S}_{22}^{+}\mathbf{S}_{21}), \tag{276}$$

Scanned with OKEN Scanner

unde $\mathbf{S}$ este matricea de covarianțe a $(\mathbf{t}_1, \mathbf{t}_2)$ partiționate în modul obișnuit ($\mathbf{S}$ este bazată pe $\{\mathbf{t}_1(X_i), \mathbf{t}_2(Y_i)\}$ $(i = 1, 2, \ldots, n)$).

TEOREMA 17 (Jupp, Mardia (1980)). i) *Parametrul* ρ^2 *este funcție de* $\boldsymbol{\Sigma}$

ii) ρ^2 *este invariant la rotații și reflecții în cazul cercului sau sferei, la transformările liniare nesingulare ale spațiilor vectoriale, la multiplicare la stînga și dreapta prin matrici ortogonale în* $O(n)$ *și* $O(p)$ *în cazul subspațiului Stiefel* $V_n(R^p)$ *și la izometrii ale subspațiului compact Riemannian.*

iii) *Dacă* X *și* Y *sînt independente atunci* $\rho^2 = 0$. *Cînd* $\boldsymbol{\Sigma}_{11}$ *și* $\boldsymbol{\Sigma}_{22}$ *au rang maxim,* $\rho^2 = 0$ *dacă numai dacă* $\boldsymbol{\Sigma}_{12} = \mathbf{0}$, *adică* $\mathbf{t}_1(X)$ *și* $\mathbf{t}_2(Y)$ *sînt necorelate.*

iv) $\rho^2 \leqslant c = \min\{\text{rang}(\boldsymbol{\Sigma}_{11}), \text{rang}(\boldsymbol{\Sigma}_{22})\}$.

v) $\rho^2 = c$ *dacă și numai dacă*

$$\mathbf{A}\mathbf{t}_1(X) + \mathbf{B}\mathbf{t}_2(Y) + \mathbf{a} = \mathbf{0}$$

cu probabilitatea unu pentru matricile $\mathbf{A}$ *și* $\mathbf{B}$ *nu amîndouă singulare și un vector* $\mathbf{a}$.

vi) *Dacă* X *și* Y *sînt independente, atunci, asimptotic* $nr^2 \sim \chi^2(\nu)$, *unde* $\nu = (\text{rang } \boldsymbol{\Sigma}_{11})(\text{rang } \boldsymbol{\Sigma}_{22})$.

Demonstrație. Din (271) și (272) urmează proprietățile i)—v). Pentru a demonstra vi), mai întîi notăm că

$$\mathbf{S}_{12} = \bar{U} - \bar{t}_1\bar{t}_2',$$

unde $\mathbf{U} = \mathbf{t}_1\mathbf{t}_2'$ și $\bar{U}$, $\bar{t}_1$, $\bar{t}_2$ reprezintă mediile de selecție. Din proprietatea de invarianță ii), putem transforma $\mathbf{t}_i$ în $\mathbf{t}_i - M(\mathbf{t}_i)$ și fără a pierde din generalitate putem presupune $M(\mathbf{t}_i) = \mathbf{0}$, $i = 1, 2$. Din legea slabă a numerelor mari (aplicată statisticii $\bar{t}_2$) și din teorma limită centrală (aplicată statisticii $\bar{t}_1$) obținem că $n^{1/2}\bar{t}_1\bar{t}_2'$ tinde în repartiție către zero. Astfel $n^{1/2}\mathbf{S}_{12}$ are asimptotic aceeași repartiție ca $n^{1/2}\bar{U}$ care este normală multidimensională cu vectorul valoare medie zero și matricea de covarianțe

$$\boldsymbol{\Sigma}_{11} \otimes \boldsymbol{\Sigma}_{22} \text{ unde } (\boldsymbol{\Sigma}_{11} \otimes \boldsymbol{\Sigma}_{22})_{ij,kl} = (\boldsymbol{\Sigma}_{11})_{ik}(\boldsymbol{\Sigma}_{22})_{jl}.$$

În consecință $nUr(\boldsymbol{\Sigma}_{11}^{+}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{+}\boldsymbol{\Sigma}_{21})$ este asimptotic repartizată ca $\chi^2(\nu)$. Din legea slabă a numerelor mari urmează că, asimptotic, nr^2 are aceeași repartiție.

Observație. Dacă repartițiile marginale aparțin unei familii exponențiale corespunzătoare atunci $\rho_2^2 = \rho^2$ și în cazul de independență nr_2^2 este asimptotic echivaint cu nr^2 și astfel, asimptotic, $nr_2^2 \sim \chi^2(pq)$.

Observație. Repartiția (nulă) asimptotică a statisticii nr_3^2 depinde de repartițiile marginale ale lui $\mathbf{t}_1$ și $\mathbf{t}_2$. Dacă aceste repartiții posedă o axă de simetrie, atunci $\rho_3^2 = \rho^2$ și în cazul de independență, nr_3^2 este asimptotic echivalent cu nr^2 și astfel, asimptotic, $nr_3^2 \sim \chi^2(pq)$.

Observație. Nu se obțin ușor formule pentru ρ^2 în termeni de parametrii în cazul familiilor generale de repartiții bidimensionale direcționale. Totuși pentru anumite familii exponențiale ρ^2 se exprimă simplu în funcție de parametrii de dependență.

Scanned with OKEN Scanner

Are loc :

TEOREMA 18. (Jupp, Mardia (1980)). *Cosiderăm familia de densităţi exponenţiale generalizate*

$$\exp\{\alpha(\mathbf{A}, \mathbf{t}_1) + \mathbf{t}_1'\mathbf{A}\mathbf{t}_2 + \beta(\mathbf{A}, \mathbf{t}_2) - \mathfrak{K}(\mathbf{A})\}$$

în raport cu orice repartiţie astfel încît X şi Y sînt independende pe M şi N. Fie Σ_{11} şi Σ_{22} matricile de covarianţă ale lui $\mathbf{t}_1$ şi $\mathbf{t}_2$ pentru repartiţia cu $\mathbf{A} = 0$. Atunci

$$\rho^2 = \mathrm{Ur}(\Sigma_{11}\mathbf{A}\Sigma_{22}\mathbf{A}') + \mathbf{0}(\|\mathbf{A}\|^3), \tag{277}$$

unde $\|\mathbf{A}\|^2 = \mathrm{Ur}(\mathbf{A}\mathbf{A}')$.

Demonstraţie. Notăm prin $M_{\mathbf{A}}(\cdot)$, $D^2_{\mathbf{A}}(\cdot)$ şi $\mathrm{cov}_{\mathbf{A}}(\cdot, \cdot)$ mediile, dispersiile şi covarianţele pentru densitatea de parametru $\mathbf{A}$ şi definim $\Sigma_{12}(\mathbf{A}) = \mathrm{cov}_{\mathbf{A}}(\mathbf{t}_1, \mathbf{t}_2)$. Deoarece (277) este invariant la translaţia lui $\mathbf{t}_i$ la $\mathbf{t}_i - M_0(\mathbf{t}_i)$, pentru simplificarea calculelor putem presupune fără pierdere de generalitate că $\mathbf{M}_0(t_i) = \mathbf{0}$, $i = 1, 2$. Dacă v este o funcţie vectorială de X şi Y, atunci

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{A}} M_{\mathbf{A}}(v) = \mathrm{cov}_{\mathbf{A}}\left(v, \frac{\partial\alpha}{\partial\mathbf{A}} + \mathbf{t}_1\mathbf{t}_2' + \frac{\partial\beta}{\partial\mathbf{A}}\right).$$

Aplicînd această relaţie pentru $v = \mathbf{t}_1$, $\mathbf{t}_2$ şi $\mathbf{t}_1\mathbf{t}_2'$, folosind faptul că

$$\Sigma_{12}(\mathbf{A}) = M_{\mathbf{A}}(\mathbf{t}_1\mathbf{t}_2') - M_{\mathbf{A}}(\mathbf{t}_1)M_{\mathbf{A}}(\mathbf{t}_2')$$

şi luînd $\mathbf{A} = \mathbf{0}$, obţinem

$$\left[\frac{\mathrm{d}}{\mathrm{d}\mathbf{A}}\Sigma_{12}\right]_{\mathbf{A}=} = \Sigma_{11} \otimes \Sigma_{22} + \mathrm{cov}_0\left(\mathbf{t}_1\mathbf{t}_2', \frac{\partial\alpha}{\partial\mathbf{A}} + \frac{\partial\beta}{\partial\mathbf{A}}\right).$$

Deoarece $\partial\alpha/\partial\mathbf{A}$ este o funcţie de X însă nu de Y, $\mathrm{cov}\,(\mathbf{t}_1\mathbf{t}_2', \partial\alpha/\partial\mathbf{A}) = 0$. Similar, avem $\mathrm{cov}_0(\mathbf{t}_1\mathbf{t}_2', \partial\beta/\partial\mathbf{A}) = 0$ şi astfel

$$\left[\frac{\mathrm{d}}{\mathrm{d}\mathbf{A}}\Sigma_{12}\right]_{\mathbf{A}=0} = \Sigma_{11} \otimes \Sigma_{22}.$$

Deoarece $\Sigma_{12}(0) = 0$, dezvoltarea Taylor a statisticii ρ^2 dă

$$\rho^2 = \mathrm{Ur}[\{\Sigma_{11} + \mathbf{0}(\|\mathbf{A}\|)\}^{-1}\{\Sigma_{11}\mathbf{A}\Sigma_{22} + \mathbf{0}(\|\mathbf{A}\|^2\} \times$$
$$\times \{\Sigma_{22} + \mathbf{0}(\|\mathbf{A}\|\}^{-1}\{\Sigma_{11}\mathbf{A}\Sigma_{22} + \mathbf{0}(\|A\|^2\}'], \tag{278}$$

şi urmează (277).

Pentru a ilustra acest rezultat considerăm următoarele exemple.

Exemplul 1. Pentru familia de repartiţii normale bidimensionale cu repartiţiile marginale cu mediile μ_i, dispersiile σ_i^2 $(i = 1, 2)$ şi covarianţa

Scanned with OKEN Scanner

$\alpha\sigma_1\sigma_2$, dacă luăm

$$t_1(X) = X - \mu_1,\ t_2(Y) = Y - \mu_2 \text{ și } A = \alpha\{\sigma_1\sigma_2(1 - \alpha^2)\}^{-1}$$

atunci (277) se reduce la

$$\rho^2 = \left(\frac{\alpha}{1 - \alpha^2}\right)^2 + \mathbf{0}(\alpha^3).$$

În cazul de față $\rho^2 \equiv \alpha^2$.

Exemplul 2. Pentru familia în care variabilele unghiulare $\boldsymbol{\theta}' = (\theta_1, \ldots, \theta_p)$ și variabilele liniare $\mathbf{X}' = (X_1, \ldots, X_q)$ au densitatea comună proporțională cu (Johnson, Wehrly (1978))

$$\exp\left\{-\frac{1}{2}\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\lambda}'\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{a}(\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}\mathbf{x}\right\},$$

unde

$$\mathbf{a}(\boldsymbol{\theta})' = \{a_1(\boldsymbol{\theta}), \ldots, a_q(\boldsymbol{\theta})\},\quad a_i(\boldsymbol{\theta}) = \sum_{j=1}^{p}\sum_{k=1}^{n}\{\alpha_{ijk}\cos(k\theta_j) + \beta_{ijk}\sin(k\theta_j)\},$$

avem:

$$\rho^2 = \frac{1}{2}\sum_{i,l=1}^{q}\sum_{j=1}^{p}\sum_{k=1}^{n}(\alpha_{ijk}\alpha_{ljk} + \beta_{ijk}\beta_{ljk})(\boldsymbol{\Sigma}^{-1})_{il} + \mathbf{0}(\alpha^3, \beta^3).$$

Exemplul 3. Pentru $\mathbf{x}$ și $\mathbf{y}$ pe sfera $(p - 1)$ dimensională, considerăm subfamilia cu densitățile (Mardia (1975))

$$C(\lambda, \varkappa)\exp(\varkappa_1\boldsymbol{\mu}_1'\mathbf{x} + \varkappa_{12}\mathbf{x}'\mathbf{R}\mathbf{y} + \varkappa_2\boldsymbol{\mu}_2'\mathbf{y}),$$

unde $\boldsymbol{\mu}_1$ și $\boldsymbol{\mu}_2$ sînt vectorii unitate și $\mathbf{R}$ este o matrice ortogonală. Constanta de normalizare (dată în termeni de funcții Bessel și Gegenbauer) este

$$\{C(\lambda, \varkappa)\}^{-1} = 2^{3s+2}\pi^{2s+2}\Gamma(s)\sum_{m=0}^{\infty}(s + m)\frac{I_{s+m}(\varkappa_1)I_{s+m}(\varkappa_{12})I_{s+m}(\varkappa_2)}{\varkappa_1^s\varkappa_{12}^s\varkappa_2^s}\cdot C_m^s(\cos\psi),$$

unde $s = \frac{1}{2}p - 1$ și $\cos\psi = \boldsymbol{\mu}_1'\mathbf{R}\boldsymbol{\mu}_2$,

Teorema 18 dă

$$\rho^2 = C\varkappa_{12}^2 + \mathbf{0}(\varkappa_{12}^3);$$

unde

$$C = \{a_1(\varkappa_1)a_1(\varkappa_2) + a_2(\varkappa_1)a_2(\varkappa_2)\}\cos^2\psi + \{a_1(\varkappa_1)a_2(\varkappa_2) +$$

$$+ a_2(\varkappa_1)a_1(\varkappa_2)\}\sin^2\psi + (p - 2)a_2(\varkappa_1)a_2(\varkappa_2),$$

$$a_1(\varkappa) = 1 - (p - 1)\varkappa^{-1}A(\varkappa) - \{A(\varkappa)\}^2,$$

$$a_2(\varkappa) = \varkappa^{-1}A(\varkappa),\ A(\varkappa) = I_{s+1}(\varkappa)/I_s(\varkappa).$$

Scanned with OKEN Scanner

§ 15. Corelația rangurilor

15.1. Coeficientul lui Spearman

Fie $(X_1, Y_1), \ldots, (X_n, Y_n)$ o selecție asupra vectorului aleator (X, Y). Dacă presupunem că variabila aleatoare $X(Y)$ are funcția de repartiție $F(G)$ continuă atunci variabila aleatoare $\beta = F(X)[\delta = G(Y)]$ este repartizată uniform pe $[0, 1]$. Urmează că

$$M(\beta) = M(\delta) = 1/2.$$
$$D^2(\beta) = D^2(\delta) = 1/12.$$

DEFINIȚIA 29. Coeficientul de corelație r al variabilelor β și δ poartă numele de coeficientul teoretic de corelație a rangurilor variabilelor X și Y.
Avem

$$R = r(\beta, \delta) = \frac{M\left(\beta - \frac{1}{2}\right)\left(\delta - \frac{1}{2}\right)}{1/12}. \tag{279}$$

Coeficientul de corelație a rangurilor variabilelor X și Y de selecție se definește în felul următor : fie X_{ik} și Y_{ik}, $1 \leqslant i, k \leqslant n$ variabile aleatoare definite prin relațiile :

$$X_{ik} = \begin{cases} 1 & \text{dacă } X_i < X_k \\ 0 & \text{dacă } X_i = X_k \\ -1 & \text{dacă } X_i > X_k \end{cases}, \quad Y_{ik} = \begin{cases} 1 & \text{dacă } Y_i < Y_k \\ 0 & \text{dacă } Y_i = Y_k \\ -1 & \text{dacă } Y_i > Y_k \end{cases} \tag{280}$$

Fie

$$\beta_i = \sum_{k=1}^{n} X_{ik}, \quad \delta_i = \sum_{k=1}^{n} Y_{ik}. \tag{281}$$

Observație. $\beta_i(\delta_i)$ este diferența dintre numărul variabilelor $X_1, \ldots, \ldots, X_n (Y_1, \ldots, Y_n)$ cu valori mai mari decît valoarea variabilei $X_i(Y_i)$ și numărul variabilelor $X_1, \ldots, X_n(Y_1, \ldots, Y_n)$ cu valori mai mici decît variabila $X_i(Y_i)$

DEFINIȚIA 30 (Spearman (1904)). Raportul

$$R = \frac{\sum_{i=1}^{n} \beta_i \delta_i}{a}, \quad a = \frac{n(n-1)(n+1)}{3} \tag{282}$$

poartă numele de coeficientul de corelație a rangurilor de selecție .

Valoarea a este valoarea maximă pe care o pot avea sumele $\sum_1^n \beta_i^2$ și $\sum_1^n \delta_i^2$ și anume : $(n-1)^2 + (n-3)^2 + \ldots + (-n+1)^2 = \frac{n(n-1)(n+1)}{3}$.

Scanned with OKEN Scanner

PROPOZIȚIA 64. *Cu probabilitatea unu, variabila aleatoare R este egală cu*

$$1 - \frac{\sum_{i=1}^{n} (\beta_i - \delta_i)^2}{2a} \quad \textit{sau cu} \quad -1 + \frac{\sum_{i=1}^{n} (\beta_i + \delta_i)^2}{2a}.$$

COROLARUL 2. $-1 \leqslant R \leqslant 1$, $R = 1$ *dacă* $\beta_i - \delta_i = 0$; $1 \leqslant i \leqslant n$ *și* $R = -1$ *dacă* $\beta_i + \delta_i = 0$, $1 \leqslant i \leqslant n$.

TEOREMA 19. *Dacă* (X, Y) *este un vector normal bidimensional cu densitatea*

$$f(x, y) = \frac{1}{2\pi}\sqrt{1 - \rho^2}\exp\left[-\frac{1}{2}(x^2 - 2\rho xy + y^2)\right]$$

atunci

$$M(R) = \frac{6}{\pi} \cdot \frac{n-2}{n-1} \arcsin\frac{\rho}{2} + \frac{6}{\pi(n+1)} \arcsin \rho. \tag{283}$$

Demonstrație. Avem

$$aM(R) = \sum_i \sum_j \sum_l M(X_{ik} Y_{il}).$$

Toți termenii sumei pentru care $i = k$ sau $i = l$ sînt nuli. Prin urmare suma conține $n(n-1)(n-2)$ termeni egali în care $k \neq l$ și $n(n-1)$ termeni egali pentru care $k = l$. Urmează că

$$aM(R) = n(n-1)(n-2)M(X_{12}Y_{13}) + n(n-1)M(X_{12}\,Y_{12}),$$

de unde

$$M(R) = 3\frac{n-2}{n+1} M(X_{12}Y_{13}) + \frac{3}{n+1} M(X_{12}Y_{12}). \tag{284}$$

Variabila aleatoare $X_{12}\,Y_{12}$ ia valoarea 1 dacă ($X_1 < X_2$ și $Y_1 < Y_2$) sau dacă ($X_1 > X_2$ și $Y_1 > Y_2$) și -1 dacă ($X_1 < X_2$ și $Y_1 > Y_2$) sau dacă ($X_1 > X_2$ și $Y_1 < Y_2$). Probabilitatea inegalităților $X_1 < X_2$ și $Y_1 < Y_2$ este

$$p_1 \equiv \frac{1}{(2\pi)^2}(1-\rho^2) \iiiint\limits_{x_1 < x_2,\ y_1 < y_2} \exp\left[-\frac{1}{2}(x_1^2 - 2\rho x_1 y_1 + y_1^2)\right] \times$$

$$\times \exp\left[-\frac{1}{2}(x_2^2 - 2\rho x_2 y_2 + y_2^2)\right] dx_1 dx_2 dy_1 dy_2 = \frac{1}{2\pi}\arccos(-\rho).$$

Aceeași valoare o are și probabilitatea p_2 a inegalităților $X_1 > X_2$ și $Y_1 > Y_2$. Probabilitatea inegalităților $X_1 < X_2$ și $Y_1 > Y_2$

$$p_3 = \frac{1}{2\pi}\arccos\rho$$

este egală cu probabilitatea p_4 a inegalităților $X_1 > X_2$ și $Y_1 < Y_2$.

Scanned with OKEN Scanner

Urmează că

$$M(X_{12}Y_{12}) = (p_1 + p_2)1 + (p_3 + p_4)(-1) = 2p_1 - 2p_2 =$$

$$= \frac{1}{\pi} \arccos(-\rho) - \frac{1}{\pi} \arccos \rho,$$

sau

$$(285)\quad M(X_{12}Y_{12}) = \frac{1}{\pi}\left(\frac{\pi}{2} + \arcsin \rho\right) - \frac{1}{\pi}\left(\frac{\pi}{2} - \arcsin \rho\right) = \frac{2}{\pi} \arcsin \rho.$$

Variabila aleatoare $X_{12}Y_{13}$ ia valoarea 1 dacă ($X_1 < X_2$ și $Y_1 < Y_3$) sau dacă ($X_1 > X_2$ și $Y_1 > Y_3$) cu probabilitatea

$$\frac{1}{2\pi} \arccos \frac{-\rho}{2}$$

și valoarea -1 dacă ($X_1 < X_2$ și $Y_1 > Y_3$) sau dacă ($X_1 > X_2$ și $Y_1 < Y_3$) cu probabilitatea

$$\frac{1}{2\pi} \arccos \frac{\rho}{2}.$$

Urmează că

$$(286)\quad M(X_{12}Y_{13}) = \frac{1}{\pi} \arccos \frac{-\rho}{2} + \frac{1}{\pi} \arccos \frac{\rho}{2} = \frac{2}{\pi} \arcsin \frac{\rho}{2}.$$

Ținînd seama în (284) de (285) și (286) teorema este demonstrată.

Observație. Pentru valori mari ale lui n, (283) se poate scrie

$$(287)\quad M(R) \simeq \frac{6}{\pi} \arcsin \frac{\rho}{2}$$

$$(288)\quad \rho \simeq 2 \sin \frac{\pi}{6} \cdot M(R).$$

TEOREMA 20. *Dacă X și Y (normale sau nu) sînt independente, atunci*

i) $M(R) = 0$ *și*

ii) $D^2(R) = 1/(n-1)$.

Demonstrație i) În acest caz variabilele β_i și δ_i sînt și ele independente și cum

$$M(\beta_i) = M(\delta_i) = 0 \text{ urmează că } M(R) = 0.$$

ii) Deoarece putem scrie

$$a^2R^2 = \sum_{i=1}^{n} \beta_i\delta_i \sum_{k=1}^{n} \beta_k\delta_k = \sum_{i,k=1}^{n} \beta_i\beta_k\delta_i\delta_k,$$

Scanned with OKEN Scanner

urmează că

$$a^2D^2(R) = a^2M(R^2) = \sum_{i,k=1}^{\bullet} M(\beta_i\beta_k)\,M(\delta_i\delta_k).$$

Suma anterioară conține n termeni egali în care $i = k$ și $n(n-1)$ termeni egali în care $i \neq k$. Avem:

$$a^2D^2(R) = nM(\beta_1^2)M(\delta_1^2) + n(n-1)M(\beta_1\beta_2)M(\delta_1\delta_2).$$

Deoarece $\sum_1^n \beta_i = 0$, urmează că

$$0 = M\left(\beta_1 \sum_{i=1}^{n} \beta_i\right) = M(\beta_1^2) + (n-1)M(\beta_1\beta_2),$$

de unde

$$M(\beta_1\beta_2) = -\frac{1}{n-1}M(\beta_1^2).$$

Analog, obținem:

$$M(\delta_1\delta_2) = -\frac{1}{n-1}M(\delta_1^2).$$

Deoarece $P\left(\sum_1^n \beta_i^2 = a\right) = 1$, obținem $\sum_1^n M(\beta_1^2) = nM(\beta_1^2) = a$ sau $M(\beta_1^2) = \frac{a}{n}$ și analog, $M(\delta_1^2) = \frac{a}{n}$.

Urmează că

$$M(\beta_1\beta_2) = M(\delta_1\delta_2) = -\frac{n}{n+1},$$

ceea ce demonstrează ii).

TEOREMA 21. *Avem*

i) $M(R^{2k-1}) = 0$ *si*

ii) $\lim_{n\to\infty} (n-1)^k M(R^{2k}) = \frac{(2k)!}{2^k k!}$, $k = 1, 2, \ldots$

COROLARUL 3. *Asimptotic*

$$\sqrt{n-1}\,R \sim N(0, 1).$$

Demonstrația rezultă din faptul că $(2k)!/2^k k!$ este momentul de ordin $2k$ al repartiției $N(0, 1)$.

Scanned with OKEN Scanner

15.2. Coeficientul T al lui Kendall

DEFINIȚIA 31 (Kendall 1948)). Raportul

$$T = \frac{\sum_{i=1}^{n} \sum_{k=1}^{n} X_{ik} Y_{ik}}{n(n-1)} \tag{289}$$

poartă numele de coeficientul de corelație al rangurilor de selecție.

Valoarea maximă a sumei de la numărător este $n(n-1)$, iar valoarea minimă $-n(n-1)$, ceea ce arată că $-1 \leqslant T \leqslant 1$, $T = -1$ dacă $X_{ik} + Y_{ik} = 0$ și $T = 1$ dacă $X_{ik} - Y_{ik} = 0$, $1 \leqslant i, k \leqslant n$.

TEOREMA 22. *Dacă X și Y sînt variabile aleatoare independente (normale sau nu), atunci*

$$M(T) = 0$$

$$D^2(T) = \frac{1}{n(n-1)} M\left(\sum_{\substack{i,k=1 \\ i<k}}^{n} X_{ik} Y_{ik} \right)^2 = \frac{2(2n+5)}{9n(n-1)}.$$

TEOREMA 23. *Avem*

$$M(T^{2k-1}) = 0$$

$$\lim_{n \to \infty} \left[\frac{9n(n-1)}{2(2n+5)} \right]^k M(T^{2k}) = \frac{(2k)!}{2^k k!}, \quad k = 1, 2, \ldots$$

COROLARUL 4. *Asimptotic*

$$\sqrt{\frac{9n(n-1)}{2(2n+5)}} \cdot T \sim N(0,1).$$

TEOREMA 24. *Dacă (X, Y) are o repartiție normală bidimensională de densitatea* (1), *atunci*

$$M(T) = \frac{2}{\pi} \arcsin \rho. \tag{290}$$

Demonstrația rezultă dacă în

$$M(T) = \frac{1}{n(n-1)} M\left(\sum_{i,k=1}^{n} X_{ik} Y_{ik} \right) = \frac{1}{n(n-1)} n(n-1) M(X_{12} Y_{12})$$

ținem seama de (285).

Observație. Din (290), găsim

$$\rho = \sin \frac{\pi}{2} \cdot M(T). \tag{291}$$

Scanned with OKEN Scanner

§ 16. Alte măsuri ale dependenței

16.1. Extinderea noțiunii de coeficient de corelație la variabile oarecare

Fie X și Y două variabile aleatoare definite pe cîmpul de probabilitate $\{\Omega, \mathcal{K}, P\}$. Putem introduce mărimea

$$r_{a,b}(X, Y) = \frac{\int_{A_a \cap B_b} X(\omega) Y(\omega) \mathrm{d}P(\omega)}{\left(\int_{A_a} |X(\omega)|^2 \mathrm{d}P(\omega)\right)^{1/2} \left(\int_{B_b} |Y(\omega)|^2 \mathrm{d}P(\omega)\right)^{1/2}} \tag{292}$$

unde

$$A_a = \{\omega \,||X(\omega)| \leqslant a\}, \quad B_b = \{\omega \,||Y(\omega)| \leqslant b\}, \quad a, b \in R^+,$$

$(A_a)_{a \in R^+}$ și $(B_b)_{b \in R^+}$ fiind familii crescătoare, ale căror limite sînt Ω.

Avem

$$|r_{a,b}(X, Y)| \leqslant 1.$$

Introducem mărimile

$$\underline{r}(X, Y) = \varliminf_{a,b \to \infty} r_{a,b}(X, Y), \quad \bar{r}(X, Y) = \varlimsup_{a,b \to \infty} r_{a,b}(X, Y).$$

Definiția 32 (Onicescu, Săcuiu (1964)). Mărimea

$$\underline{r}(X, Y) = \bar{r}(X, Y) = r(X, Y). \tag{293}$$

dacă există, poartă numele de coeficientul asimptotic de corelație.

Observație. i) $r(X, Y)$ poate să existe chiar și pentru variabile aleatoare X și Y care nu admit dispersii finite.

ii) Dacă există $D^2(X)$, $D^2(Y)$ coeficienții de corelație : clasic $\rho(X, Y)$ și asimptotic coincind.

16.2. Un coeficient de corelație global

Săcuiu (1970), referindu-se la n variabile aleatoare introduce un coeficient de corelație, care în cazul $n = 2$ se reduce la coeficientul de corelație obișnuit.

Definiția 33 (Săcuiu (1970)). Vom numi covarianța variabilelor aleatoare $X_1, X_2, \ldots, X_n$ următoarea expresie :

$$\operatorname{cov}(X_1, X_2, \ldots, X_n) = M\left(\prod_{j=1}^{n} (X_j - m_j)\right). \tag{294}$$

Scanned with OKEN Scanner

DEFINIȚIA 34. (Săcuiu (1970)). Vom numi coeficient de corelație al variabilelor aleatoare $X_1, X_2, \ldots, X_n$ următorul raport

$$r_{X_1,X_2,\ldots,X_n} = \frac{M\left(\prod_{j=1}^{n}(X_j - m_j)\right)}{\left(\prod_{j=1}^{n} M(|X_j - m_j|^n)\right)^{1/n}}. \tag{295}$$

Avem

$$\operatorname{cov}(X_1, X_2, \ldots, X_n) = \operatorname{cov}(X_{i_1}, X_{i_2}, \ldots, X_{i_n}) \tag{296}$$

$$r_{X_1,X_2,\ldots,X_n} = r_{X_{i_1},X_{i_2},\ldots,X_{i_n}} \tag{297}$$

unde $(i_1, i_2, \ldots, i_n)$ este o permutare a mulțimii $(1, 2, \ldots, n)$.

TEOREMA 25 (Săcuiu (1970)). *Dacă variabilele aleatoare $X_1, X_2, \ldots, X_n$ sînt total independente, atunci*

$$r_{X_1,X_2,\ldots,X_n} = 0.$$

TEOREMA 26 (Săcuiu (1970)). *Dacă variabilele $X_1, X_2, \ldots, X_n$ sînt independente s cîte s $(s<n)$, atunci*

$$r_{X_1,X_2,\ldots,X_n} = \frac{M\left(\prod_{j=1}^{n} X_j\right) - \sum_{j=1}^{n} m_j M\left(\prod_{k=1}^{n(j)} X_k\right) + \ldots}{\left(\prod_{j=1}^{n} M(|X_j - m_j|^n)\right)^{1/n}}$$

$$\frac{\ldots + (-1)^{n-s-1} \sum_{j_1, j_2 \ldots, j_{n-s-1}=1}^{n} m_{j_1} m_{j_2} \ldots m_{j_{n-s-1}} M\left(\prod_{k=1}^{n(j_1, j_2, \ldots, j_{n-s-1})} X_k\right) +}{\left(\prod_{j=1}^{n} M(|X_j - m_j|^n)\right)^{1/n}}$$

$$\frac{+(-1)^n (1 - C_n^1 + C_n^2 - \ldots + (-1)^s C_n^s) \prod_{j=1}^{n} m_j}{\left(\prod_{j=1}^{n} M(|X_j - m_j|^n)\right)^{1/n}}.$$

DEFINIȚIA 35 (Săcuiu 1970)). Variabilele aleatoare $X_1, X_2, \ldots, X_n$ sînt total necorelate dacă

$$\operatorname{cov}(X_1, X_2, \ldots, X_n) = 0.$$

TEOREMA 27 (Săcuiu (1970)) (r_1). *Dacă $\prod_{j=1}^{n} M(|X_j - m_j|^n) \neq 0$, atunci $r_{X_1,X_2,\ldots,X_n} = 0$ dacă și numai dacă variabilele aleatoare $X_1, X_2, \ldots, X_n$ sînt total necorelate.*

Scanned with OKEN Scanner

(r_2) $r_{\underbrace{x, x, \ldots, x}_{n}} = 1$ pentru fiecare $n = 2k$

(r_3) $|r_{x_1, x_2, \ldots, x_n}| \leqslant 1$

(r_4) $|r_{x_1, x_2, \ldots, x_n}| = 1$ *dacă şi numai dacă* $X_1, X_2, \ldots, X_n$ *satisfac următoarele egalităţi*

$$|X_j - m_j|^n = |\nu_j| |X_1 - m_1|^n, \quad 2 \leqslant j \leqslant n$$

unde ν_j *sînt, aproape sigur, constante.*

Observaţie. În (1969), Săcuiu obţine rezultate similare pentru cazul a trei vaiabile aleatoare.

16.3. Corelaţia informaţională şi coeficientul de corelaţie informaţională

16.3.1. Repartiţii discrete

DEFINIŢIA 36 (Onicescu (1972)). Fie A şi B două populaţii avînd m caracteristici comune $\gamma_1, \gamma_2, \ldots, \gamma_m$; $p_1, p_2, \ldots, p_m$ fiind probabilităţile respective pentru populaţia A şi $q_1, q_2, \ldots, q_m$ probabilităţile pentru populaţia B; aceste sisteme fiind complete în sensul că,

$$p_j \geqslant 0, \ \sum_{j=1}^{m} p_j = 1; \ q_j \geqslant 0, \ \sum_{j=1}^{m} q_j = 1.$$

Corelaţia informaţională între A şi B relativă la caracteristicile comune este dată de expresia

$$\Gamma_{AB} = \sum_{j=1}^{m} p_j q_j. \tag{298}$$

Coeficientul de corelaţie informaţională respectiv este dat de raportul,

$$\gamma_{AB} = \frac{\Gamma_{AB}}{\sqrt{\mathscr{E}_A \mathscr{E}_B}}, \tag{299}$$

unde

$$\mathscr{E}_A = \sum_{j=1}^{m} p_j^2, \quad \mathscr{E}_B = \sum_{j=1}^{m} q_j^2 \tag{300}$$

sînt energiile informaţionale ale populaţiilor respective.

TEOREMA 28 (Onicescu (1972)). $\operatorname{Max}(\min p_j, \min q_j) \leqslant \Gamma_{AB} \leqslant \min(\max p_j; \max q_j)$.

COROLARUL 5. *Cînd* $\min p_j = \min q_j = 0$ *şi* $\max p_j = \max q_j = 1$ *rezultă* $0 \leqslant \Gamma_{AB} \leqslant 1$.

Deci în general $0 \leqslant \Gamma_{AB} \leqslant 1$.

Scanned with OKEN Scanner

TEOREMA 29 (Onicescu (1972)) i). *Corelația Γ_{AB} este egală cu 1 dacă și numai dacă aceiași caracteristică γ_j de exemplu, este exclusivă în cele două populații $p_j = q_j = 1$.*

Corelația Γ_{AB} este egală cu 0 dacă și numai dacă populațiile A și B sînt într-o relație de străinătate în ceea ce privește caracteristica γ_j, adică cel puțin una din p_j, q_j este nulă pentru $j = 1, 2, \ldots, m$.

ii) $0 \leqslant \gamma_{AB} \leqslant 1$

Coeficientul de corelație informațională este mai mic decît 1 afară de cazul cînd cele două repartiții sînt identice, în care caz el are valoarea egală cu 1.

Observație. Energia informațională $\mathcal{E}_A = \sum_{j=1}^{m} p_j^2$, a populației A sau a repartiției $(p_1, p_2, \ldots, p_m)$ a caracteristicelor $\gamma_1, \gamma_2, \ldots, \gamma_m$ reprezintă corelația acestei informații cu ea însăși, sau încă media informațiilor $p_1, p_2, \ldots, p_m$ pe care le posedăm asupra caracterelor acestei repartiții. Este o valoare globală atașată acestei repartiții de aceiași natură ca și entropia lui Shannon, care este media informațiilor $\log p_j^{-1}$ $(j = 1, \ldots m)$. Considerată ca un indicator global al repartițiilor caracterelor γ, ea se bucură de proprietățile din

TEOREMA 30 (Onicescu (1966)). i) *În cazul unei uniformități totale, ceea ce este același lucru cu al unei nedeterminări totale, care corespunde probabilităților,*

$$p_1 = p_2 = \ldots = p_m = \frac{1}{m}$$

$\mathcal{E}_A$ are valoarea minimă $= \frac{1}{m}$.

ii) *Valorile lui $\mathcal{E}_A$ sînt cuprinse între $\frac{1}{m}$ și 1.*

iii) *Energia informațională scade cînd nedeterminarea sau uniformitatea crește.*

iv) *Dacă sistemul de caractere $\gamma'(\gamma'_1, \gamma'_2, \ldots, \gamma'_m)$ al populației A provine din $\gamma(\gamma_1, \ldots, \gamma_m)$ prin descompunerea caracterelor sale, energia $\mathcal{E}_A$ a sistemului se micșorează:*

(301) $$\mathcal{E}'_A \leqslant \mathcal{E}_A$$

v) *Legea de compunere a energiilor informaționale a două sisteme independente este multiplicativă.*

Observație. Posedînd aceste proprietăți, energia informațională constituie un indicator de structură a unei populații semnificativ pentru uniformitatea și concentrația sa. Ea se dovedește utilă cînd se urmărește variațiile globale obținute cînd se trece de al o populație la alta și în evoluția în timp a structurii unei populții oarecare. Ea intră deasemenea cum am văzut, ca o componentă esențială în expresia corelației celor două structuri.

Considerăm trei populații A, B, C cu repartițiile discrete p_j, q_j, $\Gamma_j (j = 1, 2, \ldots, m)$ ale caracterelor comune $\gamma(\gamma_1, \gamma_2, \ldots, \gamma_m)$.

DEFINIȚIA 37. (Onicescu (1972)). *Corelația informațională* a populațiilor A, B, C este dată de expresia *simetrică*

$$\Gamma_{ABC} = \sum_{j=1}^{m} p_j q_j \gamma_j \tag{298'}$$

și *coeficientul de corelație* informațională de

$$\gamma_{ABC} = \frac{\Gamma_{ABC}}{\sqrt[3]{F_A F_B F_C}} \tag{299'}$$

unde $F_A = \sum_1^m p_j^3$, $F_B = \sum_1^m q_j^3$, $F_C = \sum_1^m \Gamma_j^3$.

TEOREMA 31. (Onicescu (1972)) i) *γ_{ABC} este nul dacă și numai dacă $\Gamma_{ABC} = 0$, adică dacă cele trei populații sînt în stare globală de str ăinătat în raport cu caracterele γ.*

ii) $0 \leqslant \gamma_{ABC} \leqslant 1$

egalitatea avînd loc dacă și numai dacă cele trei populații au aceeași repartiție a caracterelor (sînt identice $p_j = q_j = \gamma_j$).

Se poate verifica deasemenea extinderea proprietăților energiilor informaționale expresiilor de forma $\sum_{j=1}^{n} p_k^j$ precum și verificarea afirmației relative la egalitatea cu 1 a $\gamma_{ABC\ldots L}$, adică pentru cazul a k ($k > 2$) populații.

16.3.2. Repartiții cu densitatea continuă

Fie d domeniul de variație al caracteristicii x; d poate fi finit sau infinit. Dacă $\rho_A(x)$ și $\rho_B(x)$ sînt densitățile presupuse continue ale celor două repartiții relative la populațiile A și B vom avea

$$\int_d \rho_A(x)\mathrm{d}x = 1, \quad \int_d \rho_B(x)\mathrm{d}x = 1. \tag{302}$$

DEFINIȚIA 38 (Onicescu (1972)). *Coeficientul de corelație informațională* al celor două repartiții va fi, prin definiție, limita, dacă există, a repartiției

$$\gamma_{AB}^{(\Delta)} = \frac{\sum_j \rho_A(x_j)\rho_B(x_j)(x_{j+1} - x_j)^2}{\sqrt{\left[\sum_l \rho_A^2(x_l)(x_{l+1} - x_l)^2\right]\left[\sum_k \rho_B^2(x_k)(x_{k+1} - x_k)^2\right]}} \tag{303}$$

cînd norma diviziunii $\Delta : \{x_0, x_1, x_2, \ldots, x_n\}$ a lui d, tinde către zero.

Scanned with OKEN Scanner

Fiind dată continuitatea lui $\rho_A(x)$ și $\rho_B(x)$ și cele două integrale (302) se pot întotdeauna lua diviziuni egale între ele și egale de exemplu cu $\frac{1}{n}$. În acest caz expresia lui $\gamma_{AB}^{(\Delta)}$ se reduce la,

$$\gamma_{AB}^{(\Delta)} = \frac{\sum \rho_A(x_j)\rho_B(x_j)(x_{j+1} - x_j)}{\sqrt{\left[\sum_l \rho_A^2(x_l)(x_{l+1} - x_l)\right]\left[\sum_k \rho_B^2(x_k)(x_{k+1} - x_k)\right]}} \tag{303'}$$

și dacă $n \to \infty$ se obține limita

$$\gamma_{AB} = \frac{\int_d \rho_A(x)\rho_B(x)\,dx}{\sqrt{\left[\int_d \rho_A^2(x)\,dx\right]\left[\int_d \rho_B^2(x)\,dx\right]}} \tag{303''}$$

integrale având o valoare bine definită chiar dacă $d = (-\infty, +\infty)$. Expresia (303'') poate deci să reprezinte coeficientul de corelație informațională între populațiile A și B.

Observație. Expresia acestui coeficient ne sugerează să luăm prin definiție ca expresie a *corelației informaționale* Γ_{AB} și a energiei informaționale $\mathscr{E}_A, \mathscr{E}_B$:

$$\Gamma_{AB} = \int_d \rho_A(x)\rho_B(x)\,dx \tag{304}$$

$$\mathscr{E}_A(x) = \int_d \rho_A^2(x)dx, \quad \mathscr{E}_B(x) = \int_d \rho_B^2(x)\,dx \tag{305}$$

care nu sînt altceva decît părțile principale ale extensiei ce se obține pornind de la repartițiile finite.

TEOREMA 32 (Onicescu (1972)).

$$0 \leqslant \gamma_{AB} \leqslant 1 \tag{306}$$

Dacă deplasarea caracteristicii pentru una din populații este egală cu a, presupunînd că baza caracteristicii este axa reală, atunci are loc

DEFINIȚIA 39 (Onicescu (1972)). Corelația informațională este

$$\Gamma_{AB} = \int_{-\infty}^{\infty} \rho_A(x)\rho_B(x - a)dx \tag{307}$$

și coeficientul de corelație informațională este

$$\gamma_{AB} = \frac{\Gamma_{AB}(a)}{\sqrt{\mathscr{E}_A\mathscr{E}_B}}. \tag{308}$$

Observație. Dacă $\rho_A(x) \equiv \rho_B(x) = \rho(x)$, atunci se obține corelația foarte utilă în aplicații

$$\Gamma(a) = \int \rho(x)\rho(x-a)\mathrm{d}x$$

$$\gamma(a) = \frac{\Gamma(a)}{\mathscr{E}}.$$

În cazul cînd caracterele unei populații sînt reprezentabile prin vectori euclidieni N dimensionali cu densitatea de repartiție continuă $\rho(\mathbf{x})(\rho(\mathbf{x}) \geqslant 0, \int_{E_N} \rho(\mathbf{x})\mathrm{d}\mathbf{x} = 1)$ are loc

DEFINIȚIA 40 (Onicescu 1972)) Mărimile

(309) $$\mathscr{E}_A = \int_d \rho^2(\mathbf{x})\mathrm{d}\mathbf{x}$$

(310) $$\Gamma_{AB} = \int_d \rho_A(\mathbf{x})\rho_B(\mathbf{x})\mathrm{d}\mathbf{x}$$

și

(311) $$\gamma_{AB} = \frac{\Gamma_{AB}}{\sqrt{\mathscr{E}_A\mathscr{E}_B}}$$

poartă numele de energia informațională, corelația informațională și respectiv de coeficientul de corelație informațional.

16.4. Corelație și coeficienți de corelație entropică

Pornind de la ideile lui Watanabe (1969) și Guiașu (1973) privind posibilitățile de utilizare a proprietăților entropiei lui Shannon în studiul interdependenței variabilelor aleatoare, Purcaru (1976) definește o familie de măsuri ale dependenței dintre variabilele aleatoare,sau dintre vectori aleatori, analizată apoi, atît prin prisma legăturilor acestor indicatori cu alte măsuri ale dependenței (1977, 1978, 1979), cît și în ceeace privește generalitatea și necesitatea practică a definirii și utilizării acestora (1978, 1979, 1980).

Fie $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ un vector aleator n-dimensional, ale cărui componente sînt variabile aleatoare discrete calitative sau cantitative și $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m)$, $Y_k \cap Y_l = \emptyset$, $1 \leqslant k \neq l \leqslant m$, $\bigcup_{k=1}^{m} \mathbf{Y}_k = \mathbf{X}$ o partiție a lui $\mathbf{X}$ în care $\mathbf{Y}_k$ sînt vectori aleatori n_k — dimensionali, $\sum_{k=1}^{m} n_k = n$.

Să notăm cu $H(\mathbf{X})$, $H(X_i)$ și $H(\mathbf{Y}_k)$, $1 \leqslant i \leqslant n$, $1 \leqslant k \leqslant m$, entropiile Shannon corespunzătoare vectorului aleator $\mathbf{X}$, variabilei aleatoare X_i și respectiv vectorului aleator $\mathbf{Y}_k$, definite și studiate de Guiașu ((1968), (1971), (1973), (1977)) și cu $H(\mathbf{Y}_k|\mathbf{Y}_l)$ entropia vectorului aleator $\mathbf{Y}_k$ con-

Scanned with OKEN Scanner

diționat de vectorul aleator $\mathbf{Y}_l$, $1 \leqslant k \neq l \leqslant m$, care poate avea diferite semnificații după cum $\mathbf{Y}_k$ și $\mathbf{Y}_l$ sînt unidimensionali sau multidimensionali, de asemenea studiată de Guiașu ((1968), (1973)) și Purcaru (1978). Se arată că $H(\mathbf{Y}_k|\mathbf{Y}_l) = H(\mathbf{Y}_k, \mathbf{Y}_l) - H(\mathbf{Y}_l)$. Deasemenea se poate verifica că dacă $X_i \in \mathbf{Y}_k$ atunci $H(X_i) \leqslant H(\mathbf{Y}_k) \leqslant H(\mathbf{X})$. Are loc întotdeauna inegalitatea $H(\mathbf{Y}_k|\mathbf{Y}_l) \leqslant H(\mathbf{Y}_k)$ care devine egalitate dacă și numai dacă $\mathbf{Y}_k$ și $\mathbf{Y}_l$ sînt independenți probabilistic.

Între entropia vectorului aleator $\mathbf{X}$ și entropiile vectorilor $\mathbf{Y}_k$, $1 \leqslant k \leqslant m$, în care el este partiționat, are loc relația $H(\mathbf{X}) \leqslant \sum_{k=1}^{m} H(\mathbf{Y}_k)$ cu egalitate dacă și numai dacă avem o partiționare independentă a lui $\mathbf{X}$ [Guiașu ((1968), (1973)), Purcaru ((1978), (1979))].

DEFINIȚIA 41 (Purcaru (1978)). Vom spune că vectorul aleator $\mathbf{Y}_k$ este complet dependent de vectorul aleator $\mathbf{Y}_l$ dacă $H(\mathbf{Y}_k|\mathbf{Y}_l) = 0$.

DEFINIȚIA 42 (Purcaru (1978)). Vom spune că vectorii aleatori $\mathbf{Y}_k$ și $\mathbf{Y}_l$ sînt reciproc complet dependenți dacă $H(\mathbf{Y}_k|\mathbf{Y}_l) = H(\mathbf{Y}_l|\mathbf{Y}_k) = 0$.

Ținînd cont de proprietățile entropiei (Guiașu (1968), (1971), (1973), (1977)), Purcaru ((1976), (1978), (1979), (1980)), Watanabe (1969)) să considerăm mărimea:

$$W[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] = \sum_{k=1}^{m} H(\mathbf{Y}_k) - H(\mathbf{X})$$

care reprezintă diferența dintre cantitatea de informație $\sum_{k=1}^{m} H(\mathbf{Y}_k)$ pe care o conține vectorul aleator $\mathbf{X}$ în ipoteza independenței părților sale $\mathbf{Y}_1$, $\mathbf{Y}_2, \ldots, \mathbf{Y}_m$ și cantitatea de informație $H(\mathbf{X})$ conținută de același vector aleator în ipoteza dependenței probabiliste a părților sale sau, altfel spus, diferența dintre suma entropiilor părților și entropia ansamblului.

DEFINIȚIA 43 (Purcaru (1978), (1979), (1980)). Mărimea $W[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m]$ se numește corelație entropică globală a vectorilor aleatori $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$.

Semnificația corelației globale a vectorilor $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ este aceea de forță sau intensitate a coeziunii sau interdependenței dintre aceste mărimi aleatoare.

TEOREMA 33 (Purcaru (1978), ((1979)). *Corelația entropică globală* $W[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m]$ *satisface următoarele proprietăți:*

1) *este definită pentru orice partiție* $(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m)$ *a lui* $\mathbf{X}$

2) *este simetrică în variabilele* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ *adică*

$$W[\mathbf{X}: \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] = W[\mathbf{X}; \mathbf{Y}_{k_1}, \mathbf{Y}_{k_2}, \ldots, \mathbf{Y}_{k_m}]$$

unde $\{k_1, k_2, \ldots, k_m\}$ *este o permutare arbitrară a numerelor* $\{1, 2, \ldots, m\}$.

3) $0 \leqslant W[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] < \infty$

4) $W[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] = 0$ *dacă și numai dacă vectorii aleatori* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ *sînt independenți probabilistic.*

5) *este invariantă la orice transformare biunivocă a vectorilor* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ *adică* $W[\mathbf{X}; \varphi_1(\mathbf{Y}_1), \varphi_2(\mathbf{Y}_2), \ldots, \varphi_m(\mathbf{Y}_m)] = W[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m]$ *oricare ar fi funcțiile biunivoce* $\varphi_1, \varphi_2, \ldots, \varphi_m$; $R \to R$.

Scanned with OKEN Scanner

6) $W[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] \leqslant \sum_{j=2}^{m} H(\mathbf{Y}_{k_j})$ *cu egalitate dacă și numai dacă vectorul aleator* $\mathbf{Y}_{k_j}$ *este complet dependent de familia de vectori* $\{\mathbf{Y}_{k_1}, \mathbf{Y}_{k_2} \ldots, \mathbf{Y}_{k_{j-1}}\}$ *unde* $\{k_1, k_2, \ldots, k_m\}$ *este o permutare a indicilor* $\{1, 2, \ldots, m\}$.

7) $W[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] = \sum_{j=1}^{m-1} H(\mathbf{Y}_{k_j}) = (m-1)H(\mathbf{X})$ *dacă și numai dacă vectorii aleatori* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ *sînt reciproc complet dependenți.*

8) $W[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] \leqslant \frac{m-1}{m} \sum_{k=1}^{m} H(\mathbf{Y}_k)$ *cu egalitate dacă și numai dacă vectorii aleatori* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ *sînt reciproc complet dependenți.*

9) $W[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] \leqslant \min_{1 \leqslant j \leqslant m} \left\{ \sum_{k_j=1}^{m-1} H(\mathbf{Y}_{k_j}) \right\} \leqslant \max_{1 \leqslant j \leqslant m} \left\{ \sum_{k_j=1}^{m-1} H(\mathbf{Y}_{k_j}) \right\}$

cu primă egalitate dacă și numai dacă familia de vectori aleatori pentru care s-a atins minimul este complet dependentă de celălalt vector și cu egalitate peste tot dacă și numai dacă toate cantitățile sînt reciproc complet dependente.

Ca urmare a acestei teoreme se poate deduce un rezultat important în studiul corelației entropice.

TEOREMA 34 (Purcaru (1978), (1979)). *Corelația entropică globală satisface întotdeauna următoarele inegalități:*

$$0 \leqslant W[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] \leqslant \min_{1 \leqslant j \leqslant m} \left\{ \sum_{k_j=1}^{m-1} H(\mathbf{Y}_{k_j}) \right\} \leqslant \frac{m-1}{m} \sum_{k=1}^{m} H(\mathbf{Y}_k) \leqslant$$

$$\leqslant \max_{1 \leqslant j \leqslant m} \left\{ \sum_{k_j=1}^{m-1} H(\mathbf{Y}_{k_j}) \right\} \leqslant (m-1)H(\mathbf{X})$$

egalitatea cu zero avînd loc dacă și numai dacă vectorii aleatori $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ *sînt independenți, iar celelalte egalități avînd loc simultan dacă și numai dacă vectorii aleatori* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ *sînt reciproc complet dependenți.*

Deși comodă în utilizare, corelația entropică globală are o scară mare de valori în aprecierea gradului de dependență globală dintre vectorii aleatori $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$.

Ca urmare a acestui fapt și ținînd cont de proprietățile entropice globale, Purcaru normează această măsură introducînd noi indicatori ai dependenței globale ce iau valori în intervalul [0, 1]. Să notăm astfel cu (Purcaru (1978), (1979))

$$G_1[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] = \frac{W[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m]}{(m-1)H(\mathbf{X})};$$

$$G_2[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] = \frac{W[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m]}{\max_{1 \leqslant j \leqslant m} \left\{ \sum_{k_j=1}^{m-1} H(\mathbf{Y}_{k_j}) \right\}};$$

$$G_3[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] = \frac{W[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m]}{\frac{m-1}{m} \sum_{k=1}^{m} H(\mathbf{Y}_k)};$$

$$G_4[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] = \frac{W[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m]}{\min_{1 \leqslant j \leqslant m} \left\{ \sum_{k_j=1}^{m-1} H(\mathbf{Y}_{k_j}) \right\}}.$$

Scanned with OKEN Scanner

DEFINIȚIA 44 (Purcaru (1978), (1979)). *Mărimile* $G_i[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m]$, $1 \leqslant i \leqslant 4$, *se numesc coeficienți de corelație entropică globală a vectorilor aleatori* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$.

TEOREMA 35 (Purcaru (1978), (1979)). *Coeficienții de corelație entropică globală satisfac următoarele:*

1) *sînt definiți pentru orice vectori aleatori* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ *neconstanți cu probabilitatea unu,*

2) *sînt simetrici ca funcții de variabile* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ *adică*

$$G_i[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] = G_i[\mathbf{X}; \mathbf{Y}_{k_1}, \mathbf{Y}_{k_2}, \ldots, \mathbf{Y}_{k_m}], \quad 1 \leqslant i \leqslant 4$$

pentru orice permutare $\{k_1, k_2, \ldots, k_m\}$ *a numerelor* $\{1, 2, \ldots, m\}$,

3) $0 \leqslant G_i[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] \leqslant G_{i+1}[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m]$, $1 \leqslant i \leqslant 3$,

4) $G_i[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] = 0$ *dacă și numai dacă vectorii aleatori* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ *sînt probabilistic independenți,* $1 \leqslant i \leqslant 4$,

5) $G_i[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] = 1$, $1 \leqslant i \leqslant 4$, *dacă și numai dacă vectorii aleatori* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ *sînt reciproc complet dependenți,*

6) *sînt invarianți la orice transformare biunivocă a vectorilor aleatori* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$.

O proprietate importantă a coeficienților de corelație entropică globală constă în aceea că ei definesc de fapt o întreagă familie de măsuri entropice ale dependenței globale.

TEOREMA 36 (Purcaru (1978), (1979)). *Orice combinație liniară convexă a coeficienților de corelație entropică globală* $G_i[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m]$,

$$G_\alpha[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] = \sum_{i=1}^{4} \alpha_i G_i[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m], \quad 0 \leqslant \alpha_i \leqslant 1, \sum_{i=1}^{4} \alpha_i = 1$$

este de asemenea coeficient de corelație entropică globală a vectorilor aleatori $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$.

DEFINIȚIA 45. Vom spune că vectorii aleatori $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ sînt global necorelați dacă $G_i[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] = 0$, $1 \leqslant i \leqslant 4$.

DEFINIȚIA 46. Vom spune că vectorii aleatori $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ sînt perfect global corelați dacă $G_i[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m] = 1$, $1 \leqslant i \leqslant 4$.

Trebuie remarcat că definițiile de mai sus nu impun satisfacerea simultană a egalității cu zero și respectiv cu unu pentru cei patru coeficienți, întrucît acestea sînt îndeplinite simultan ca urmare a proprietăților corelației entropice globale. Menționăm de asemenea că în practică poate fi ales spre utilizare oricare din cei patru coeficienți de corelație entropică globală și chiar oricare din elementele familiei $G_\alpha[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m]$ generată de aceștia, toți coincizînd în situațiile extreme ale necorelării globale sau corelării globale perfecte.

Analizînd familia de măsuri entropice generată de corelația entropică globală $W[\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m]$, constatăm că (Purcaru (1978), (1979))

1) dacă $m = n$, avem indicatorii corelației globale a variabilelor aleatoare $x_1, x_2, \ldots, x_n$

2) dacă $m = 2$, $n_1 \geqslant 2$, și $n_2 \geqslant 2$, avem indicatorii corelației entropice dintre doi vectori aleatori

Scanned with OKEN Scanner

3) dacă $m = 2$, $n_1 = 1$ și $n_2 = n - 1$, avem indicatorii corelației entropice dintre o variabilă aleatoare și un vector aleator

4) dacă $m = n = 2$ avem indicatorii corelației entropice dintre două variabile aleatoare,

formulele de mai sus rămînînd valabile peste tot dar adaptate în scriere corespunzător fiecărui caz.

O analiză sumară a familiei de măsuri entropice ale interdependenței scoate în evidență două calități definitorii ale acestora și anume :

a) generalitata aplicabilității măsurilor entropice ale dependenței ca urmare a faptului că ele nu sînt definite decît cu ajutorul repartițiilor probabiliste ale vectorilor aleatori considerați, ceea ce le face utilizabile atît pentru vectori aleatori cu componente calitative cît și pentru cei cu componente cantitative.

b) elasticitatea în utilizare a măsurilor entropice ale dependenței ca urmare a ușurinței cu care se poate trece de la două la mai multe variabile aleatoare, sau, la doi sau mai mulți vectori aleatori, fără introducerea unor condiții suplimentare cînd se face extinderea măsurilor.

Ca urmare a acestor calități, corelația entropică poate fi considerată ca un generator al unei familii de măsuri ale dependenței probabiliste.

Scanned with OKEN Scanner

Capitolul XXIX

INFERENȚE PRIVIND COEFICIENȚII DE CORELAȚIE ȘI MATRICILE DE COVARIANȚE

§ 1. Estimarea parametrilor repartiției normale bidimensionale

1.1. Selecție totală (nedesfăcută)

Fie (X_i, Y_i), $i = 1, 2, \ldots, n$ o selecție asupra unui vector normal bidimensional (X, Y) cu densitatea:

$$(1) \qquad \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right\}.$$

Pentru funcția de verosimilitate, avem

$$\ln P = -n\ln(2\pi) - \frac{n}{2}\ln(1-\rho^2) - \frac{1}{2(1-\rho^2)}\sum(x^2 - 2\rho xy + y^2)$$

și ecuația de verosimilitate este

$$(2) \quad \frac{\partial \ln P}{\partial \rho} = \frac{n\rho}{1-\rho^2} - \frac{\rho}{(1-\rho^2)^2}\left(\sum x_i^2 - 2\rho\sum x_i y_i + \sum y_i^2\right) + \frac{1}{1-\rho^2}\sum x_i y_i = 0.$$

(2) se mai poate scrie

$$(2') \quad \rho(1-\rho^2) + (1+\rho^2)\frac{1}{n}\sum x_i y_i - \rho\left(\frac{1}{n}\sum x_i^2 + \frac{1}{n}\sum y_i^2\right) = 0.$$

(2') are trei rădăcini, două dintre ele putînd fi complexe. Dacă cele trei rădăcini sînt reale și dau valori ale lui ρ în domeniul admisibil, atunci vom alege ca estimator de V.M. aceea valoare care face maximă funcția de verosimilitate.

Dacă exprimăm ecuația cubică în forma

$$(2'') \qquad \lambda^3 + p\lambda + q = 0$$

cu

$$(3) \qquad \lambda = \rho - \frac{1}{3n}\sum x_i y_i,$$

Scanned with OKEN Scanner

condiţia de a avea numai o rădăcină reală este ca

$$4p^3 + 27q^2 > 0 \tag{4}$$

şi este sigur realizată cînd $p > 0$, unde :

$$p = \frac{1}{n}\sum x_i^2 + \frac{1}{n}\sum y_i^2 - \frac{1}{3}\left(\frac{1}{n}\sum x_i y_i\right)^2 - 1. \tag{5}$$

Deoarece momentele de selecţie sînt estimatori consistenţi pentru momentele teoretice corespunzătoare, din (5), p converge în probabilitate către

$$\left(1 + 1 - \frac{1}{3}\rho^2 - 1\right) = 1 - \frac{1}{3}\rho^2 > 0.$$

Astfel, pentru selecţii de volum mare, ecuaţia de verosimilităţi are o singură rădăcină reală şi aceasta este estimatorul de V.M., rădăcinile complexe avînd valori inadmisibile.

PROPOZIŢIA 1. *Asimptotic, estimatorul de verosimilitate maximă* $\hat{\rho}$ *are dispersia*

$$D^2(\hat{\rho}) = \frac{(1-\rho^2)^2}{n(1+\rho^2)}. \tag{6}$$

Demonstraţie. Derivînd (2) în raport cu ρ obţinem :

$$\frac{\partial^2 \ln P}{\partial \rho^2} = \frac{n(1+\rho^2)}{(1-\rho^2)^2} - \frac{(1+3\rho^2)}{(1-\rho^2)^3}\left(\sum x_i^2 - 2\rho\sum x_i y_i + \sum y_i^2\right) +$$

$$+ \frac{4\rho}{(1-\rho^2)^2}\sum x_i y_i.$$

Ţinînd seama că

$$M(X^2) = M(Y^2) = 1 \quad \text{şi} \quad M(XY) = \rho,$$

găsim

$$M\left(\frac{\partial^2 \ln P}{\partial \rho^2}\right) = \frac{n(1+\rho^2)}{(1-\rho^2)^2} - \frac{2n(1+3\rho^2)}{(1-\rho^2)^2} + \frac{4n\rho^2}{(1-\rho^2)^2} = -\frac{n(1+\rho^2)}{(1-\rho^2)^2}. \tag{7}$$

Propoziţia este demonstrată dacă ţinem seama de teorema IX.7.

Observaţie. În cazul în care repartiţia depinde de k parametrii, putem fi interesaţi în estimarea oricărui număr dintre ei de la 1 la k, ceilalţi fiind cunoscuţi. Estimatorii de verosimilitate maximă pentru parametrii consideraţi vor fi obţinuţi prin alegerea din cele k ecuaţii de verosimilitate (IX.29), a submulţimii de ecuaţii corespunzătoare parametrilor consideraţi şi din rezolvarea lor. Prin acest procedeu, estimatorii de V.M vor fi afectaţi de cunoaşterea celorlalţi parametrii ai repartiţiei.

Scanned with OKEN Scanner

Dacă (X_i, Y_i), $i = 1, 2, \ldots, n$ este o selecţie asupra vectorului (X, Y) normal bidimensional de densitate (XXVIII.1), atunci funcţia de verosimilitate este

$$P(\mathbf{x}, \mathbf{y} \mid \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2})^n} \times$$

(8)

$$\times \exp\left\{-\frac{1}{2(1-\rho^2)} \sum\left[\left(\frac{x_i-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{(x_i-\mu_1)(y_i-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{y_i-\mu_2}{\sigma_2}\right)^2\right]\right.$$

Din (8) obţinem

(9)
$$\begin{cases} \dfrac{\partial \ln P}{\partial \mu_1} = \dfrac{n}{\sigma_1(1-\rho^2)}\left[\dfrac{\bar{x}-\mu_1}{\sigma_1} - \rho\dfrac{\bar{y}-\mu_2}{\sigma_2}\right] \\ \dfrac{\partial \ln P}{\partial \mu_2} = \dfrac{n}{\sigma_2(1-\rho^2)}\left[\dfrac{\bar{y}-\mu_2}{\sigma_2} - \rho\dfrac{\bar{x}-\mu_1}{\sigma_1}\right] \end{cases}$$

(10)
$$\begin{cases} \dfrac{\partial \ln P}{\partial \sigma_1^2} = -\dfrac{1}{2\sigma_1^2(1-\rho^2)}\left[n(1-\rho^2) - \dfrac{\sum(x_i-\mu_1)^2}{\sigma_1^2} + \rho\dfrac{\sum(x_i-\mu_1)(y_i-\mu_2)}{\sigma_1\sigma_2}\right] \\ \dfrac{\partial \ln P}{\partial \sigma_2^2} = -\dfrac{1}{2\sigma_2^2(1-\rho^2)}\left[n(1-\rho^2) - \dfrac{\sum(y_i-\mu_2)^2}{\sigma_2^2} + \rho\dfrac{\sum(x_i-\mu_1)(y_i-\mu_2)}{\sigma_1\sigma_2}\right] \end{cases}$$

(11)
$$\frac{\partial \ln P}{\partial \rho} = \frac{1}{1-\rho^2}\left\{n\rho - \frac{1}{(1-\rho^2)}\left[\rho\left(\frac{\sum(x_i-\mu_1)^2}{\sigma_1^2} + \frac{\sum(y_i-\mu_2)^2}{\sigma_2^2}\right) - \right.\right.$$
$$\left.\left. - (1+\rho^2)\frac{\sum(x_i-\mu_1)(y_i-\mu_2)}{\sigma_1\sigma_2}\right]\right\}.$$

a) Presupunînd că μ_1, μ_2, σ_1, σ_2 sînt cunoscuţi, pentru a estima pe ρ rezolvăm (11). Se obţine pentru estimatorul de verosimilitate maximă al parametrului ρ o ecuaţie cubică.

b) Presupunînd că dorim să estimă σ_1^2, σ_2^2 şi ρ, μ_1 şi μ_2 fiind cunoscuţi. Avem de rezolvat trei ecuaţii de verosimilitate (10) şi (11). Obţinem

(12)
$$\begin{cases} n(1-\rho^2) = \dfrac{\sum(x_i-\mu_1)^2}{\sigma_1^2} - \rho\dfrac{\sum(x_i-\mu_1)(y_i-\mu_2)}{\sigma_1\sigma_2} \\ n(1-\rho^2) = \dfrac{\sum(y_i-\mu_2)^2}{\sigma_2^2} - \rho\dfrac{\sum(x_i-\mu_1)(y_i-\mu_2)}{\sigma_1\sigma_2} \end{cases}$$

şi

(13)
$$n(1-\rho^2) = \frac{\sum(x_i-\mu_1)^2}{\sigma_1^2} + \frac{\sum(y_i-\mu_2)^2}{\sigma_2^2} - \frac{1+\rho^2}{\rho}\cdot\frac{\sum(x_i-\mu_1)(y_i-\mu_2)}{\sigma_1\sigma_2}.$$

Scanned with OKEN Scanner

Scăzînd (13) din suma ecuațiilor din (12), obținem

$$n(1-\rho^2)=\frac{1-\rho^2}{\rho}\cdot\frac{\sum(x_i-\mu_1)(y_i-\mu_2)}{\sigma_1\sigma_2}$$

sau

$$\rho=\frac{\frac{1}{n}\sum(x_i-\mu_1)(y_i-\mu_2)}{\sigma_1\sigma_2}. \tag{14}$$

Ținînd seama de (14) în (12), obținem

$$\begin{cases}\hat{\sigma}_1^2=\dfrac{1}{n}\sum(x_i-\mu_1)^2\\ \\ \hat{\sigma}_2^2=\dfrac{1}{n}\sum(y_i-\mu_2)^2\end{cases} \tag{15}$$

și din (14)

$$\hat{\rho}=\frac{\frac{1}{n}\sum(x_i-\mu_1)(y_i-\mu_2)}{\hat{\sigma}_1\hat{\sigma}_2}. \tag{16}$$

Am demonstrat

PROPOZIȚIA 2. *În cazul în care μ_1 și μ_2 sînt cunoscuți estimatorul de verosimilitate maximă pentru ρ este dat de* (16).

Are loc [illegible]

PROPOZIȚIA 3. *$(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\rho})$ constituie o mulțime suficientă de estimatori pentru $(\sigma_1^2, \sigma_2^2, \rho)$.*

PROPOZIȚIA 4. *Pentru selecții de volum mare, matricea de covariațe a estimatorilor $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ și $\hat{\rho}$ este*

$$\Sigma_3=\frac{1}{n}\begin{pmatrix}2\sigma_1^4 & 2\rho^2\sigma_1^2\sigma_2^2 & \rho(1-\rho^2)\sigma_1^2\\ 2\rho^2\sigma_1^2\sigma_2^2 & 2\sigma_2^4 & \rho(1-\rho^2)\sigma_2^2\\ \rho(1-\rho^2)\sigma_1^2 & \rho(1-\rho^2)\sigma_2^2 & (1-\rho^2)^2\end{pmatrix} \tag{17}$$

Demonstrație. Deoarece $(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\rho})$ sînt suficiente pentru $(\sigma_1^2, \sigma_2^2, \rho)$ (propoziția 3), din teorema IX.10.

$$\Sigma_3^{-1}=(\Sigma_{rs}^{-1})=-\left(\frac{\partial^2\ln P}{\partial\theta_r\partial\theta_s}\right)_{\hat{\theta}=\theta}=\frac{n}{1-\rho^2}\begin{pmatrix}\frac{2-\rho^2}{4\sigma_1^4} & \frac{-\rho^2}{4\sigma_1^2\sigma_2^2} & \frac{-\rho}{2\sigma_1^2}\\ \frac{-\rho}{4\sigma_1^2\sigma_2^2} & \frac{2-\rho^2}{4\sigma_2^4} & \frac{-\rho}{2\sigma_2^2}\\ \frac{-\rho}{2\sigma_1^2} & \frac{-\rho}{2\sigma_2^2} & \frac{1+\rho^2}{1-\rho^2}\end{pmatrix}. \tag{18}$$

Inversînd (18), propoziția este demonstrată.

Scanned with OKEN Scanner

c) Dacă dorim să estimăm toți parametrii repartiției, rezolvăm (9), (10) și (11) împreună. (9) se reduce la

$$(19)\qquad \begin{cases} \dfrac{\bar{x}-\mu_1}{\sigma_1} = \rho\dfrac{\bar{y}-\mu_2}{\sigma_2} \\ \dfrac{\bar{y}-\mu_2}{\sigma_2} = \rho\dfrac{\bar{x}-\mu_1}{\sigma_1} \end{cases}$$

o pereche de ecuații a cărei singură soluție este

$$(20)\qquad \bar{x}-\mu_1 = \bar{y}-\mu_2 = 0$$

(15), (16) și (20) dau estimatorii de verosimilitate maximă

$$(21)\qquad \begin{cases} \hat{\mu}_1 = \bar{x}, \quad \hat{\sigma}_1^2 = \dfrac{1}{n}\sum (x_i-\bar{x})^2 \\ \hat{\mu}_2 = \bar{y}, \quad \hat{\sigma}_2^2 = \dfrac{1}{n}\sum (y_i-\bar{y})^2 \end{cases}, \qquad \hat{\rho} = \frac{\dfrac{1}{n}\sum (x_i-\bar{x})(y_i-\bar{y})}{\hat{\sigma}_1\hat{\sigma}_2}.$$

Am demonstrat

Propoziția 5. *Estimatorii de verosimilitate maximă pentru parametrii* $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ *și* ρ *sînt momentele de selecție corespunzătoare.*

Are loc

Propoziția 6. (21) *constituie o mulțime suficientă de estimatori pentru* $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

Observație. Matricea Σ_3^{-1} din (18) este o parte a inversei matricii Σ_5^{-1}. Scriind parametrii în ordinea $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$, (18) va fi minorul principal inferior al lui Σ_5^{-1}. Din (9) obținem

$$(22)\qquad \frac{\partial^2 \ln P}{\partial \mu_1^2} = \frac{-n}{\sigma_1^2(1-\rho^2)}, \quad \frac{\partial^2 \ln P}{\partial \mu_1 \partial \mu_2} = \frac{n\rho}{\sigma_1\sigma_2(1-\rho^2)}, \quad \frac{\partial^2 \ln P}{\partial \mu_2^2} = \frac{-n}{\sigma_2^2(1-\rho^2)},$$

în timp ce

$$(23)\qquad \frac{\partial^2 \ln P}{\partial \mu_i \partial \sigma_1^2} = \frac{\partial^2 \ln P}{\partial \mu_i \partial \sigma_2^2} = \frac{\partial^2 \ln P}{\partial \mu_i \partial \rho} = 0, \quad i = 1, 2$$

în $\bar{x} = \mu_1$, $\bar{y} = \mu_2$. Inversa matricei de covariante a estimatorilor de verosimilitate maximă pentru μ_1 și μ_2 este

$$(24)\qquad \Sigma_2^{-1} = \frac{n}{(1-\rho^2)} \begin{pmatrix} \dfrac{1}{\sigma_1^2} & \dfrac{-\rho}{\sigma\,\sigma_2} \\ \dfrac{-\rho}{\sigma_1\sigma_2} & \dfrac{1}{\sigma_2^2} \end{pmatrix}.$$

Scanned with OKEN Scanner

Putem scrie

(25) $$\Sigma_5^{-1} = \begin{pmatrix} \Sigma_2^{-1} & 0 \\ 0 & \Sigma_3^{-1} \end{pmatrix}$$

şi cum

(26) $$\Sigma_2 = \frac{1}{n}\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

obţinem

(27) $$\Sigma_5 = \begin{pmatrix} \Sigma_2 & 0 \\ 0 & \Sigma_3 \end{pmatrix}$$

Am redemonstrat teorema III.12 şi anume „în selecţii asupra unui vector normal bidimensional mediile de selecţie sînt independente de dispersiile şi covarianţele de selecţie" şi am demonstrat

PROPOZIŢIA 7. *Corelaţia dintre mediile de selecţie este* ρ *şi corelaţia dintre dispersiile de selecţie este* ρ^2.

1.2. Estimarea coeficientului de corelaţie dintr-o selecţie aleatoare desfăcută

1.2.1. Estimarea parametrilor necunoscuţi

Presupunem că avem o selecţie de volum n asupra vectorului (X, Y) normal bidimensional de densitate (XXVIII.1). De asemenea presupunem că înainte ca valorile de selecţie să fie înregistrate, fiecare vector (X_i, Y_i) este desfăcut în două componente separate X_i şi Y_i şi aceste observaţii sînt posibile numai în forma

(28) $$X_1, X_2, \ldots, X_n \quad \text{şi} \quad U_1, U_2, \ldots, U_n,$$

unde $U_1, U_2, \ldots, U_n$ este o permutare necunoscută a observaţiilor $Y_1, Y_2 \ldots, Y_n$. Deoarece perechile din selecţia originală nu sînt cunoscute putem da

DEFINIŢIA 1 (DeGroot, Feder, Goel (1971)). Valorile observate (28) poartă numele de selecţie aleatoare desfăcută.

Fără a pierde din generalitate presupunem că toate valorile observate sînt distincte. Notăm prin Φ mulţimea tuturor permutărilor $\phi = (\phi(1), \ldots, \phi(n))$ ale întregilor $1, 2, \ldots, n$. Împerecherea valorilor observate în selecţia desfăcută în acord cu permutarea ϕ, înseamnă împerecherea lui X_i cu $U_{\phi(i)}$ pentru $i = 1, \ldots, n$.

Vom determina estimatorii de verosimilitate maximă ai parametrilor $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ şi ρ, şi a permutării necunoscute ϕ care specifică împerecherile în selecţia iniţială. Exceptînd o constantă logaritmul funcţiei de verosimilitate este

$$\mathscr{P}(\phi, \rho, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2 | \mathbf{x}, \mathbf{u}) = -\frac{n}{2}\ln(1-\rho^2) - \frac{n}{2}\ln\sigma_1^2 - \frac{n}{2}\ln\sigma_2^2 -$$

(29)

$$-\frac{1}{2(1-\rho^2)}\left\{\sum_1^n\left(\frac{x_i-\mu_1}{\sigma_1}\right)^2 + \sum_1^n\left(\frac{y_i-\mu_2}{\sigma_2}\right)^2 - 2\rho\sum_1^n\frac{(x_i-\mu_1)}{\sigma_1}\frac{(u_{\phi(i)}-\mu_2)}{\sigma_2}\right\}.$$

Scanned with OKEN Scanner

Pentru fiecare ϕ fixat, folosind selecţia cu perechile în acord cu ϕ, logaritmul funcţiei de verosimilitate este maximizat cînd :

$$\hat{\mu}_1 = \bar{x}, \quad \hat{\mu}_2 = \bar{u}, \quad \hat{\sigma}_1^2 = m_{2x}, \quad \hat{\sigma}^2 = m_{2u} \text{ şi } \hat{\rho} = r_\phi$$

unde

$$r_\phi = \frac{1}{n} \sum_1^n (x_i - \bar{x}) u_{\phi(i)} / (\hat{\sigma}_1 \hat{\sigma}_2). \tag{30}$$

Atunci logaritmul funcţiei de verosimilitate (29) evaluat în aceste puncte este

$$\hat{\mathscr{P}}(\phi) = \mathscr{P}(\phi, r_\phi, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2 | \mathbf{x}, \mathbf{u}) = -\frac{n}{2} \log (1 - r_\phi^2) - n. \tag{31}$$

Trebuie să determinăm maximul acestei funcţii după toate permutările ϕ.

Fie

$$r_{\min} = \min_{\phi \in \Phi} r_\phi \text{ şi } r_{\max} = \max_{\phi \in \Phi} r_\phi \tag{31}$$

şi fie $x_{(1)} < \ldots < x_{(n)}$ şi $u_{(1)} < \ldots < u_{(n)}$ statisticile de ordine corespunzătoare x-lor, respectiv u-lor.

PROPOZIŢIA 8. *$r_{\max}$ este obţinut din permutarea ϕ^0 în care $x_{(i)}$ este împerecheat cu $u_{(i)}$, iar $r_{\min}$ este obţinut din permutarea ϕ_0 în care $x_{(i)}$ este împerecheat cu $u_{(n+1-i)}$.*

PROPOZIŢIA 9. *Avem* $\sum_{\phi \in \Phi} r_\phi = 0$.

Observaţie. Deoarece valoarea lui r_ϕ nu este aceeaşi pentru fiecare permutare $\phi \in \Phi$,

$$r_{\min} < 0 \text{ şi } r_{\max} > 0.$$

PROPOZIŢIA 10. (DeGroot, Goel (1980)). *Estimatorii de verosimilitate maximă pentru ρ şi ϕ sînt daţi de*

$$\begin{cases} \hat{\rho} = r_{\max} \text{ şi } \hat{\phi} = \phi^0 \text{ dacă } r_{\max} \geqslant |r_{\min}|, \\ \hat{\rho} = r_{\min} \text{ şi } \hat{\phi} = \phi_0 \text{ dacă } r_{\max} \leqslant |r_{\min}| \end{cases} \tag{33}$$

Presupunem că μ_1 şi μ_2 sînt cunoscute şi fără a pierde din generalitate presupunem că $\mu_1 = \mu_2 = 0$. Pentru $\phi \in \Phi$, fie

$$r_\phi^* = \left(\sum_1^n x_i u_{\phi(i)} \right) \Big/ \left[\left(\sum_1^n x_i^2 \right) \left(\sum_1^n u_i^2 \right) \right]^{1/2} \tag{34}$$

şi

$$r_{\min}^* = \min_{\phi \in \Phi} r_\phi^* \text{ şi } r_{\max}^* = \max_{\phi \in \Phi} r_\phi^*. \tag{35}$$

PROPOZIŢIA 11. *Estimatorii de verosimilitate maximă pentru ρ şi ϕ sînt daţi de*

$$\begin{cases} \hat{\rho} = r_{\max}^* \text{ şi } \hat{\phi} = \phi^0 \text{ dacă } r_{\max}^* \geqslant |r_{\min}^*| \\ \hat{\rho} = r_{\min}^* \text{ şi } \hat{\phi} = \phi_0 \text{ dacă } r_{\max}^* \leqslant |r_{\min}^*|. \end{cases} \tag{36}$$

Scanned with OKEN Scanner

Observație. În acest caz nu este necesar că $r^*_{max} \geqslant 0$ și $r^*_{min} \leqslant 0$. Totuși, din (36),

$$\text{dacă } r^*_{max} < 0, \text{ atunci } \hat{\rho} = r^*_{min} \text{ și}$$

$$\text{dacă } r^*_{min} > 0, \text{ atunci } \hat{\rho} = r^*_{max}.$$

Cînd μ_1, μ_2, σ_1^2 și σ_2^2 sînt cunoscuți, nu obținem o expresie simplă pentru $\hat{\rho}$. Fără a pierde din generalitate luăm

$$\mu_1 = \mu_2 = 0 \quad \text{și} \quad \sigma_1^2 = \sigma_2^2 = 1.$$

Maximizînd funcția $\mathscr{P}$ dată de (29) după $\phi \in \Phi$ pentru orice valoare fixată a lui ρ și derivînd în raport cu ρ, obținem că $\hat{\rho}$ trebuie să fie o soluție a ecuației (similară cu (2′)).

$$(37) \qquad n\rho(1-\rho^2) - \rho \sum_1^n (x_i^2 + u_i^2) + (1+\rho^2) \sum_1^n x_i u_{\phi(i)} = 0,$$

unde $\phi = \phi^0$ pentru $\rho > 0$ și $\phi = \phi_0$ pentru $\rho < 0$.

1.2.2. Funcția de verosimilitate integrată

Pentru fiecare $\phi \in \Phi$, fie $g(\mathbf{x}, \mathbf{u} \mid \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho, \phi)$ densitatea de repartiție comună a observațiilor $X_1, X_2, \ldots, X_n$ și $U_1, U_2, \ldots, U_n$ cînd ele sînt împerecheate în acord cu permutarea ϕ. Vom presupune că permutarea adevărată necunoscută are o repartiție apriorică. În particular, presupunem că apriori cele $n!$ permutări din Φ sînt egal probabile în ceea ce privește generarea selecției inițiale. Prin urmare, densitatea de repartiție comună a observațiilor $X_1, X_2, \ldots, X_n$ și $U_1, U_2, \ldots, U_n$ în selecția desfăcută, pentru valori date ale parametrilor μ_1, μ_2, σ_1^2, σ_2^2 și ρ, este media

$$(38) \qquad \frac{1}{n!} \sum_{\phi \in \Phi} g(\mathbf{x}, \mathbf{u} \mid \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho, \phi).$$

Ca funcție de parametrii μ_1, μ_2, σ_1^2, σ_2^2 și ρ pentru valori date pentru $\mathbf{x}$ și $\mathbf{u}$, (38) este verosimilitatea integrată definită în IX.10.2. Deoarece ϕ este discretă, DeGroot, Goel (1980) numesc funcția (38), funcția de verosimilitate sumată. Cum (38) nu conține parametrul nedorit ϕ, această funcție este folositoare în probleme în care sîntem interesați în inferențe asupra parametrilor rămași. Unii statisticieni privesc ϕ ca o dată lipsă și nu ca parametru. În aceste cazuri, se folosește algoritmul dat de Dempster, Laird, Rubin (1977). În cele ce urmează pornind de la (38) folosim acest algoritm, pentru obținerea estimațiilor.

Exceptînd o constantă, logaritmul funcției de verosimilitate este

$$\mathscr{P}(\rho, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid \mathbf{x}, \mathbf{u}) = -\frac{n}{2} \ln [(1-\rho^2)\sigma_1^2\sigma_2^2] -$$

$$(39) \qquad -\frac{1}{2(1-\rho^2)} \sum_1^n \left[\left(\frac{x_i - \mu_1}{\sigma_1} \right)^2 + \left(\frac{u_i - \mu_2}{\sigma_2} \right)^2 \right] +$$

$$+ \ln \sum_{\phi \in \Phi} \exp \left[\frac{\rho}{1-\rho^2} \sum_1^n \left(\frac{x_i - \mu_1}{\sigma_1} \right) \left(\frac{u_{\phi(i)} - \mu_2}{\sigma_2} \right) \right].$$

Scanned with OKEN Scanner

Ca și în cazul unei selecții nedesfăcute, estimatorii de verosimilitate maximă pentru μ_1, μ_2, σ_1^2 și σ_2^2 sînt $\bar{x}$, $\bar{y}$, m_{2x}, m_{2y}.

Logaritmul funcției de verosimilitate evaluat în $\mu_1 = \hat{\mu}_1$, $\mu_2 = \hat{\mu}_2$, $\sigma_1^2 = \hat{\sigma}_1^2$ și $\sigma_2^2 = \hat{\sigma}_2^2$ este (omițînd termenii care nu conțin pe ρ)

$$\hat{\mathscr{P}}(\rho) = \mathscr{P}(\rho, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2 \mid \mathbf{x}, \mathbf{u}) =$$

$$= -\frac{n}{2}\ln(1-\rho^2) - \frac{n}{1-\rho^2} + \ln \sum_{\phi \in \Phi} \exp\left(\frac{n\rho r_\phi}{1-\rho^2}\right),$$

unde r_ϕ este definit în (30). Estimatorul de verosimilitate maximă pentru ρ va fi soluția ecuației

$$\frac{d\hat{\mathscr{P}}(\rho)}{d\rho} = 0. \tag{41}$$

Vom scrie (41) sub o altă formă.

Pentru $\phi \in \Phi$ și pentru orice valoare a lui $\rho(|\rho| < 1)$, fie

$$\xi(\phi \mid \rho) = \frac{\exp\left(\frac{n\rho r_\phi}{1-\rho^2}\right)}{\sum_{\psi \in \Phi} \exp\left(\frac{n\rho r_\psi}{1-\rho^2}\right)}. \tag{42}$$

Astfel, pentru fiecare valoare dată a lui ρ, $\xi(\phi \mid \rho)$ reprezintă o repartiție pe cele $n!$ permutări ale lui Φ. Derivînd (40), ecuația (41) poate fi scrisă sub forma

$$\sum_{\phi \in \Phi} r_\phi \xi(\phi \mid \rho) = \rho. \tag{43}$$

PROPOZIȚIA 12. *Pentru orice valoare a lui ρ care satisface (43) relația $d^2\hat{\mathscr{P}}(\rho)/d\rho^2 < 0$ va fi satisfăcută dacă și numai dacă*

$$\sum_{\phi \in \Phi} r_\phi^2 \xi(\phi \mid \rho) < \rho^2 + \frac{(1-\rho^2)^2}{n(1+\rho^2)}. \tag{44}$$

Deci estimatorul de verosimilitate maximă $\hat{\rho}$ va fi valoarea lui ρ ce satisface (43) și (44). Valoarea lui $\hat{\rho}$ poate fi determinată iterativ, însă pentru valori ale lui n moderat de mari, costul acestui calcul poate fi foarte mare. Totuși din (43) și (44) rezultă

TEOREMA 1 (DeGroot, Goel (1980)). i) *Funcția $\hat{\mathscr{P}}(\rho)$ este crescătoare pentru $\rho \leqslant r_{\min}$ și descrescătoare pentru $\rho \geqslant r_{\max}$.*

ii) *Pentru valoarea $\rho = 0$, $\hat{\mathscr{P}}(\rho)$ are un minim local*

iii) *În fiecare din intervalele*

$$r_{\min} < \rho < 0 \text{ și } 0 < \rho < r_{\max}$$

există cel puțin un maxim local pentru $\hat{\mathscr{P}}(\rho)$.

Scanned with OKEN Scanner

Determinăm

$$i_n(\rho) = M\left[-\frac{\partial^2 \mathscr{P}(\rho, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2 | \mathbf{X}, \mathbf{U})}{\partial \rho^2} \right] \tag{45}$$

unde $\mathbf{X} = (X_1, \ldots, X_n)$ și $\mathbf{U} = (U_1, \ldots, U_n)$ sînt variabile aleatoare în selecția desfăcută și $\mathscr{P}$ este dat de (39).

Pentru orice variabile $X_1, \ldots, X_n$ și $U_1, \ldots, U_n$ fie

$$\beta_\phi = \frac{1}{\sigma_1 \sigma_2} \sum_i (X_i - \mu_1)(U_{\phi(i)} - \mu_2) \tag{46}$$

și

$$\alpha_\phi = \frac{\exp\left(\frac{\rho \beta_\phi}{1 - \rho^2}\right)}{\sum_{\psi \in \Phi} \exp\left(\frac{\rho \beta_\psi}{1 - \rho^2}\right)}. \tag{47}$$

Din (39) obținem

$$i_n(\rho) = -\frac{n(1 + \rho^2)}{(1 - \rho^2)^2} + \frac{2n(1 + 3\rho^2)}{(1 - \rho^2)^3} - \frac{2\rho(3 + \rho^2)}{(1 - \rho^2)^3} M\left(\sum_{\phi \in \Phi} \beta_\phi \alpha_\phi\right) -$$

$$- \frac{(1 + \rho^2)^2}{(1 - \rho^2)^4} M\left[\sum_{\phi \in \Phi} \beta_\phi^2 \alpha_\phi - \left(\sum_{\phi \in \Phi} \beta_\phi \alpha_\phi\right)^2\right], \tag{48}$$

unde mediile în (48) sînt calculate în presupunerea că cele n perechi de variabile aleatoare $(X_1, U_1), \ldots, (X_n, U_n)$ constituie o selecție asupra unui vector normal bidimensional de densitate (XXVIII.1). Observăm că valorile medii din (48) rămîn neschimbate dacă presupunem în (XXVIII.1), (46) și (47) că $\mu_1 = \mu_2$ și $\sigma_1 = \sigma_2 = 1$. Urmează că

$$M\left(\sum_{\phi \in \Phi} \beta_\phi \alpha_\phi\right) = M\left(\sum_i X_i U_i\right) = n\rho. \tag{49}$$

Fie

$$\gamma = \frac{1}{n} \cdot M\left[\sum_{\phi \in \Phi} \beta_\phi^2 \alpha_\phi - \left(\sum_{\phi \in \Phi} \beta_\phi \alpha_\phi\right)^2\right]. \tag{50}$$

Atunci, din (48), (49) și (50), găsim

$$i_n(\rho) = \frac{n(1 + \rho^2)}{(1 - \rho^2)^2} \left[1 - \frac{(1 + \rho^2)}{(1 - \rho^2)^2} \gamma\right]. \tag{51}$$

Din (47), pentru orice valori $X_1, \ldots, X_n$ și $U_1, \ldots, U_n$ date, găsim

$$\alpha_\phi > 0 \text{ pentru } \phi \in \Phi \text{ și } \sum_{\phi \in \Phi} \alpha_\phi = 1.$$

Scanned with OKEN Scanner

Deci, din (50), $\gamma > 0$. Evaluînd $M(\sum_{\phi \in \Phi} \beta_\phi^2 \alpha_\phi)$ găsim că

$$\gamma = 1 + (n+1)\rho^2 - \frac{1}{n} M[(\sum_{\phi \in \Phi} \beta_\phi z_\phi)^2]. \tag{52}$$

Fie

$$\gamma_1 = \frac{(1+\rho^2)\gamma}{(1-\rho^2)^2} \quad \text{și} \quad \gamma_2 = \frac{\rho^2 \gamma}{1-\rho^2}. \tag{53}$$

Cu acestea, matricea de covarianțe Σ_3^{-1}, bazată pe selecția desfăcută este:

$$\Sigma_3^{-1} = \frac{n}{1-\rho^2} \begin{pmatrix} \frac{1}{\sigma_1^2}(2-\rho^2-\gamma_2) & \frac{1}{\sigma_1\sigma_2}(-\rho-\gamma_2) & -\frac{\rho}{\sigma_1}(1-\gamma_1) \\ & \frac{1}{\sigma_2^2}(2-\rho^2-\gamma_2) & -\frac{\rho}{\sigma_2}(1-\gamma_1) \\ & & \frac{1+\rho^2}{1-\rho^2}(1-\gamma_1) \end{pmatrix}. \tag{54}$$

Observație. În această problemă nu avem n observații bidimensionale independente și identic repartizate. Prin urmare nu sînt satisfăcute condițiile pentru validitatea proprietăților asimptotice ale estimatorilor de verosimilitate maximă și Σ_3 nu este necesar matricea de covarianțe asimptotică a estimatorilor de verosimilitate maximă $\hat{\sigma}_1$, $\hat{\sigma}_2$ și $\hat{\rho}$.

PROPOZIȚIA 13. *Dacă* μ_1, μ_2, σ_1 *și* σ_2 *sînt cunoscuți,* $i_n(\rho)$ *(dată de* (51)) *este informația lui Fisher pentru selecția desfăcută.*

TEOREMA 2 (DeGroot, Goel (1980))

i) $i_n(0) = 1$ *pentru* $n = 1, 2, \ldots$

ii) *Pentru* $-1 < \rho < 1$ *și* $n = 1, 2, \ldots$

$$i_n(\rho) > i_1(\rho) = \frac{1+\rho^2}{(1-\rho^2)^2}. \tag{55}$$

§ 2. Inferențe privind coeficienții de corelație cu ajutorul aproximației lui Fisher

2.1. Test pentru un ρ dat

Din propoziția XXVIII.27 urmează că

$$Z = \frac{\frac{1}{2}\log\frac{1+R}{1-R} - \frac{1}{2}\log\frac{1+\rho}{1-\rho}}{\sqrt{\frac{1}{n-3}}} \sim N(0, 1) \tag{56}$$

Scanned with OKEN Scanner

și că poate fi folosită pentru verificarea ipotezei $H_0 : \rho = \rho_0$ față de una din ipotezele alternative

$$H_1 : \rho \neq \rho_0, \quad H_1' : \rho < \rho_0, \quad H'' : \rho > \rho_0.$$

Regiunile critice corespunzătoare diferitelor ipoteze alternative ipotezei $H_0 : \rho = \rho_0$ sînt trecute în tabelul 1

Tabelul 1

$H_0 : \rho = \rho_0$		
H_1	Regiunea critică	
$\rho \neq \rho_0$	(57)	$\lvert Z \rvert > z_{1-(\alpha/2)}$
$\rho < \rho_0$	(58)	$Z < z_\alpha$
$\rho > \rho_0$	(59)	$Z > z_{1-\alpha}$

unde Z este dat de (56) cu $\rho = \rho_0$.

PROPOZIȚIA 14. *Intervalul*

$$(60) \quad \left(\frac{1}{2} \log \frac{1+R}{1-R} - z_{1-(\alpha/2)} \sqrt{\frac{1}{n-3}}, \quad \frac{1}{2} \log \frac{1+R}{1-R} + z_{1-(\alpha/2)} \sqrt{\frac{1}{n-3}} \right)$$

este un interval de încredere pentru $\frac{1}{2} \log \frac{1+\rho}{1-\rho}$ *avînd coeficientul de încredere* $1 - \alpha$.

Demonstrația rezultă din faptul că statistica (56) satisface condițiile a), b) și c) din § 2 capitolul XII pentru existența intervalelor de încredere.

2.2. Test pentru egalitatea a doi coeficienți de corelație

Două selecții de volume n_1 și n_2 au condus la coeficienții de corelație de selecție r_1 și respectiv r_2. Sînt aceste valori compatibile cu ipoteza conform căreia selecțiile provin din populații avînd același coeficient de corelație? Fie

$$z_1 = \frac{1}{2} \log \frac{1+R_1}{1-R_1} \quad \text{și} \quad z_2 = \frac{1}{2} \log \frac{1+R_2}{1-R_2}.$$

Dacă $\rho_1 = \rho_2 = \rho$, atunci asimptotic, statistica

$$(61) \qquad Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \sim N(0, 1)$$

și poate fi folosită pentru verificarea ipotezei $H_0 : \rho_1 = \rho_2$ față de una din ipotezele alternative

$$H_1 : \rho_1 \neq \rho_2, \quad H_1' : \rho_1 < \rho_2, \quad H_1'' : \rho_1 > \rho_2.$$

Scanned with OKEN Scanner

Regiunile critice corespunzătoare diferitelor ipoteze alternative ipotezei $H_0 : \rho_1 = \rho_2$ sînt trecute în tabelul 2

Tabelul 2

$H_0 : \rho_1 = \rho_2$		
H_1	Regiunea critică	
$\rho_1 \neq \rho_2$	(62)	$\lvert Z \rvert > r_{1-(\alpha/2)}$
$\rho_1 < \rho_2$	(63)	$Z < z_\alpha$
$\rho_1 > \rho_2$	(64)	$Z > z_{1-\alpha}$

2.3. Test pentru omogenitatea unei mulțimi de coeficienți de corelație

În scopul verificării ipotezei conform căreia coeficienții de corelație a k populații normale bidimensionale sînt egali, adică

$$H_0 : \rho_1 = \rho_2 = \ldots = \rho_k$$

folosim statistica

$$(65)\quad T = \sum_1^k (n_i - 3)\left(\frac{1}{2}\log\frac{1+R_i}{1-R_i}\right)^2 - \frac{1}{\sum_1^k (n_i - 3)}\left[\sum_1^k (n_i - 3)\frac{1}{2}\log\frac{1+R_i}{1-R_i}\right]^2$$

care este repartizată $\chi^2(k-1)$.

Respingem ipoteza H_0, dacă

$$(66)\qquad T > \chi^2_{1-\alpha}(k-1).$$

§ 3. Teste de independență în repartiția normală bidimensională

Pentru repartiția normală bidimensională de densitate (XXVIII.1), ne propunem să verificăm ipoteza

$\rho = 0$ (X și Y sînt variabile aleatoare independente).

Funcția de verosimilitate a selecției (X_t, Y_t), $t = \overline{1, n}$ asupra acestui vector are forma exponențială (XVI.1) cu

$$U = \sum X_t Y_t, \quad T_1 = \sum X_t^2, \quad T_2 = \sum Y_t^2, \quad T_3 = \sum X_t, \quad T_4 = \sum Y_t$$

și

$$\theta = \frac{\rho}{\sigma_1\sigma_2(1-\rho^2)}, \quad v_1 = \frac{-1}{2\sigma_1^2(1-\rho^2)}, \quad v_2 = \frac{-1}{2\sigma_2^2(1-\rho^2)}$$

$$v_3 = \frac{1}{1-\rho^2}\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2\rho}{\sigma_1\sigma_2}\right), \quad v_4 = \frac{1}{1-\rho^2}\left(\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1\rho}{\sigma_1\sigma_2}\right).$$

Scanned with OKEN Scanner

Ipoteza $H : \rho \leqslant 0$ este echivalentă cu $\theta \leqslant 0$. Deoarece coeficientul de corelaţie de selecţie R (dat de XXVIII.1) nu se schimbă cînd X_t şi Y_t sînt înlocuite prin $(X_t - \mu_1)/\sigma_1$ şi $(Y_t - \mu_2)/\sigma_2$, repartiţia lui R nu depinde de μ_1, μ_2, σ_1 sau σ_2, însă depinde numai de ρ. Prin urmare, pentru $\theta = 0$, repartiţia nu depinde de ν_1, ν_2, ν_3, ν_4 şi deci după teorema XVI.2, cînd $\theta = 0$, R este independent de (T_1, T_2, T_3, T_4). Din teorema XVI.1 urmează că testul U.C.M.P. nedeplasat respinge ipoteza H cînd

$$R \geqslant C_0 \tag{67}$$

sau echivalent, cînd

$$\frac{R}{\sqrt{(1 - R^2)/(n - 2)}} > K_0. \tag{68}$$

Statistica R este liniară în U şi repartiţia sa pentru $\rho = 0$ este simetrică faţă de zero.

Prin urmare am demonstrat

TEOREMA 3. *Testul U.C.M.P. nedeplasat pentru verificarea ipotezei* $\rho = 0$ *faţă de* $\rho \neq 0$ *respinge ipoteza nulă cînd*

$$\frac{|R|}{\sqrt{(1 - R^2)/(n - 2)}} > K_1. \tag{69}$$

Observaţie. Deoarece cînd $\rho = 0$, $\sqrt{n-2}\,R/\sqrt{1 - R^2} \sim t(n - 2)$. (propoziţia XXVIII.17), constantele K_0 şi K_1 sînt date de

$$\int_{K_0}^{\infty} g_{n-2}(t)\,dt = \alpha \quad \text{şi} \quad \int_{K_1}^{\infty} g_{n-2}(t)\,dt = \frac{\alpha}{2}, \tag{70}$$

unde g_{n-2} este densitatea de repartiţie corespunzătoare unei variabile repartizate $t(n - 2)$.

Deoarece repartiţia lui R depinde numai de coeficientul de corelaţie ρ acelaşi lucru este adevărat pentru puterea acestor teste.

Samiuddin (1970), pentru a verifica ipoteze privind coeficientul de corelaţie ρ în populaţii normale bidimensionale foloseşte statistica

$$t' = \frac{r'}{\sqrt{1 - r'^2}}\sqrt{n - 2}, \tag{71}$$

unde r' poate fi obţinut, pentru r şi ρ daţi, rezolvînd

$$\frac{w - M(\rho)}{\sqrt{D^2(\rho)}} = \frac{\log(1 + r'/1 - r')}{\sqrt{D^2(0)}} \tag{72}$$

w, $M(\rho)$, $D^2(\rho)$, avînd semnificaţiile din XXVIII.§ 4.

Cînd $\rho = 0$, $t' = t$.

Folosind (71) ca $\sim t(n - 2)$, obţinem rezultate mai bune decît cu transformarea z a lui Fisher, în acelaşi mod în care expresia t dată de (XXVIII.78) pentru $\rho = 0$, dă rezultate mai bune decît transformarea z corespunzătoare. Tabela 3 confirmă această afirmaţie; ea dă funcţia de repartiţie F_1 folosind r', F_2 folosind z şi funcţia de repartiţie exactă obţinută din tabela lui David (1954).

Scanned with OKEN Scanner

În concluzie : t (definit de (XXVIII.78) are avantaje practice : dă un test bilateral corespunzător pentru ρ, are o repartiție $t(n-2)$ pentru $\rho = 0$ și orice n, sau pentru $\rho \neq 0$ și valori moderat de mari ale lui n, dă o metodă pentru obținerea unui interval de încredere pentru ρ.

Totuși t' pentru un test unilateral dă rezultate mai bune.

Tabela 3

r	F_1	F_2	Exactă
	(a) $n = 8$, $\rho = 0{,}9$		
0,99	0,99260	0,99477	0,99214
0,98	0,95826	0,95937	0,95728
0,95	0,74909	0,74344	0,75096
0,90	0,42317	0,42579	0,42572
0,80	0,13985	0,14435	0,13884
0,60	0,02439	0,02226	0,02296
0,50	0,01196	0,00949	0,01089
	(b) $n = 8$, $\rho = 0{,}3$		
0,90	0,99325	0,99520	0,99306
0,80	0,95759	0,95818	0,95765
0,70	0,89006	0,88686	0,89009
0,60	0,79796	0,79103	0,79814
0,45	0,63683	0,62930	0,63735
0,25	0,42532	0,42858	0,42518
0,00	0,21684	0,22065	0,21642
−0,20	0,10724	0,11123	0,10716
−0,35	0,05564	0,05592	0,05543
−0,60	0,01210	0,00964	0,01199

Din corolarul XXVIII.3 (XXVIII.4) rezultă un test pentru verificarea ipotezei conform căreia X și Y sînt independente, dacă volumul n al selecției este suficient de mare. Testul respinge ipoteza de independență la pragul de semnificație α dacă

(73)[(73′)] $$|\sqrt{n-1}\,R| > z_{\alpha/2}\left(\sqrt{\frac{9n(n-1)}{2(2n+5)}}\,T \sim N(0,1)\right).$$

unde

$$\Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$$

§ 4. Verificarea ipotezei de independență pentru date direcționale

În familiile din teorema XXVIII.18 independența corespunde la $\mathbf{A} = \mathbf{0}$. Pentru verificarea ipotezei de independență avem două statistici test : nr^2 și $\Lambda = -2 \log \lambda$. Are loc următorul rezultat

TEOREMA 4 (Jupp, Mardia (1980)). *Considerăm familia de densități exponențiale generalizată* :

$$\exp\{\alpha(\mathbf{A}, \mathbf{t}_1) + \mathbf{t}_1'\mathbf{A}\mathbf{t}_2 + \beta(\mathbf{A}, \mathbf{t}_2) - \varkappa(\mathbf{A})\}$$

Scanned with OKEN Scanner

în raport cu orice repartiție astfel încît X și Y sînt independente pe M și N

i) *Fie $M_0(\mathbf{t}_i)$ valoarea medie a lui $\mathbf{t}_i$ $(i = 1, 2)$ pentru repartiția cu $\mathbf{A} = \mathbf{0}$. Dacă*

$$[\partial\alpha/\partial\mathbf{A}]_{\mathbf{A}=\mathbf{0}} + \mathbf{t}_1 M_0(\mathbf{t}_2') \text{ și } [\partial\beta/\partial\mathbf{A}]_{\mathbf{A}=\mathbf{0}} + M_0(\mathbf{t}_1)\mathbf{t}_2'$$

sînt constante, atunci sub $\mathbf{A} = \mathbf{0}$ și în anumite condiții de regularitate asimptotic, Λ și nr^2 sînt echivalente.

ii) *Fie Λ^c statistica pe care se bazează testul raportului de verosimilități condiționată de $\mathbf{t}_1$ și $\mathbf{t}_2$. Dacă*

$$[\partial\alpha/\partial\mathbf{A}]_{\mathbf{A}=\mathbf{0}} \text{ și } [\partial\beta/\partial\mathbf{A}]_{\mathbf{A}=\mathbf{0}}$$

sînt funcții liniare de $\mathbf{t}_1$ și $\mathbf{t}_2$, atunci sub $\mathbf{A} = \mathbf{0}$ și în anumite condiții de regularitate, asimptotic Λ^c și nr^2 sînt echivalente.

În particular considerăm exemplele după teorema XXVIII.18 cînd numai parametrul de dependență este necunoscut.

Avem echivalența asimptotică a statisticilor Λ și nr^2 în ipoteza de independență pentru exemplele XXVIII.1, dacă $F = 0$, XXVIII.2 și XXVIII.3 dacă $\mathfrak{K}_1 = \mathfrak{K}_2 = 0$. Această echivalență asimptotică are loc de asemenea pentru familia de repartiții bidimensionale circulare cu densitățile proporționale cu

$$\exp\left[\mathfrak{K}_1 \cos\{l_1(\theta - \mu_1)\} + \mathbf{t}_1'\mathbf{A}\mathbf{t}_2 + \mathfrak{K}_2 \cos\{l_2(\phi - \mu_2)\}\right]$$

în raport cu repartițiile uniforme, pentru întregii $l_1, l_2 \geqslant 2$ și cu $\mathbf{t}_1'(\theta) = (\cos\theta, \sin\theta)$, $\mathbf{t}_2'(\phi) = (\cos\phi, \sin\phi)$. Dacă $\mathbf{A} = \mathbf{0}$, θ și ϕ au repartiții von Mises independente multimodale (Mardia (1972), Mardia, Spurr (1973)). Pentru această familie, în ipoteza de independență Λ^c și nr^2 sînt echivalente.

§ 5. Teste privind coeficientul de corelație multiplu și coeficientul de corelație parțial multiplu

5.1. Introducere

Considerăm vectorul $\mathbf{X} : 1 \times p$, partiționat ca

$$\mathbf{X} = (X_1, \mathbf{X}_2(1 \times p_2), \mathbf{X}_3(1 \times p_3))$$

având o repartiție $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ fiind partiționată ca

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{13} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23} \\ \boldsymbol{\Sigma}_{31} & \boldsymbol{\Sigma}_{32} & \boldsymbol{\Sigma}_{33} \end{pmatrix} \begin{matrix} 1 \\ p_2 \\ p_3 \end{matrix} \qquad (74)$$
$$\begin{matrix} 1 & p_2 & p_3 \end{matrix}$$

Fie $\rho_{1.23}$ și $\rho_{1.2}$ coeficienții de corelație multiplă dintre X_1 și $(\mathbf{X}_2, \mathbf{X}_3)$, și respectiv între X_1 și $\mathbf{X}_2$; aici indicii 1,2 și 3 se referă la X_1, $\mathbf{X}_2$ și res-

Scanned with OKEN Scanner

pectiv $\mathbf{X}_3$. Considerăm următoarele probleme cînd atît $\boldsymbol{\mu}$ cît și $\boldsymbol{\Sigma}$ sînt necunoscuți

i) $H_1 : \rho_{1.23} = 0$ față de $K_1 : \rho_{1.23} = \rho_{1.2} > 0$,
ii) $H_2 : \rho_{1.23} = 0$ față de $K_2 : \rho_{1.2} = 0,\ \rho_{1.23} > 0$,
iii) $H_3 : \rho_{1.23} = \rho_{1.2}$ față de $K_3 : \rho_{1.23} > \rho_{1.2}$,
iv) $H_4 : \rho_{1.2} = 0$ față de $K_4 : \rho_{1.2} > 0$.

Vom da soluțiile acestor probleme cu ajutorul a N observații independente asupra lui $\mathbf{X}$. Fie $\overline{\mathbf{X}}$ și S/N estimatorii de verosimilitate maximă pentru $\boldsymbol{\mu}$ și respectiv $\boldsymbol{\Sigma}(N > p)$.

Problemele i) — iv) sînt invariante sub grupul $\mathcal{G}$ al transformărilor g date de:

$$(75) \qquad g\mathbf{X} = (X_1 + a_1,\ \mathbf{X}_2 + \mathbf{a}_2,\ \mathbf{X}_3 + \mathbf{a}_3)\begin{pmatrix} g_{11} & 0 & 0 \\ 0 & \mathbf{G}_{22} & \mathbf{G}_{23} \\ 0 & 0 & \mathbf{G}_{33} \end{pmatrix},$$

unde $g_{11} \neq 0$, $\mathbf{G}_{22} : p_2 \times p_2$ și $\mathbf{G}_{33} : p_3 \times p_3$ sînt matrici nesingulare, $\mathbf{G}_{23}$ este $p_2 \times p_3$ și $a_1 : 1 \times 1$, $\mathbf{a}_2 : 1 \times p_2$, $\mathbf{a}_3 = 1 \times p_3$ sînt vectori reali.

5.2. Invarianți maximali și teste ale raportului de verosimilități

Transformarea indusă de g pe spațiul statisticilor suficiente $(\overline{\mathbf{X}}, \mathbf{S})$ este dată de

$$(\overline{\mathbf{X}}, \mathbf{S}) \to ((\overline{\mathbf{X}} + \mathbf{a})\mathbf{G}, \mathbf{G}'\mathbf{S}\mathbf{G}),$$

unde

$$\mathbf{a} = (a_1, a_2, a_3)$$

$$\mathbf{G} = \begin{pmatrix} g_{11} & 0 & 0 \\ 0 & \mathbf{G}_{22} & \mathbf{G}_{23} \\ 0 & 0 & \mathbf{G}_{33} \end{pmatrix}.$$

Orice funcție invariant depinde de $\overline{\mathbf{X}}$ și $\mathbf{S}$ numai prin $\mathbf{S}$. Considerăm următoarea partiție a lui $\mathbf{S}$:

$$(76) \qquad \mathbf{S} = \begin{matrix} \\ p_2 \\ p_3 \\ \\ \end{matrix}\!\!\begin{matrix} \begin{pmatrix} S_{11} & \mathbf{S}_{12} & \mathbf{S}_{13} \\ \mathbf{S}_{21} & \mathbf{S}_{22} & \mathbf{S}_{23} \\ \mathbf{S}_{31} & \mathbf{S}_{32} & \mathbf{S}_{33} \end{pmatrix} \\ \begin{matrix} 1 & p_2 & p_3 \end{matrix} \end{matrix}\!\!\begin{matrix} 1 \\ p_2 \\ p_3 \\ \\ \end{matrix}$$

Definim

$$g_{11} = S_{11}^{-1/2}, \quad \mathbf{G}_{22} = \mathbf{T}_{22}\mathbf{L}_{22}, \quad \mathbf{G}_{23} = \mathbf{T}_{23}\mathbf{L}_{33}, \quad \mathbf{G}_{33} = \mathbf{T}_{33}\mathbf{L}_{33},$$

unde

$$\mathbf{T}_{22} : p_2 \times p_2, \quad \mathbf{T}_{23} : p_2 \times p_3, \quad \mathbf{T}_{33} : p_3 \times p_3, \quad \mathbf{L}_{22} : p_2 \times p_2 \text{ și } \mathbf{L}_{33} : p_3 \times p_3$$

Scanned with OKEN Scanner

sînt definite după cum urmează

$$\mathbf{T}'\begin{pmatrix}\mathbf{S}_{22} & \mathbf{S}_{23}\\ \mathbf{S}_{32} & \mathbf{S}_{33}\end{pmatrix}\mathbf{T} = \mathbf{I}_{p_2+p_3}, \qquad \mathbf{T} = \begin{pmatrix}\mathbf{T}_{22} & \mathbf{T}_{23}\\ \mathbf{0} & \mathbf{T}_{33}\end{pmatrix},$$

$\mathbf{L}_{22}$ este o matrice ortogonală cu prima coloană proporţională cu $\mathbf{T}_{22}\mathbf{S}_{21}\mathbf{S}_{11}^{-1/2}$, $\mathbf{L}_{33}$ este o matrice ortogonală cu prima coloană proporţională cu $(\mathbf{T}'_{23}\mathbf{S}_{21} + \mathbf{T}'_{33}\mathbf{S}_{31})\mathbf{S}_{11}^{-1/2}$. Atunci

$$\mathbf{G}'\mathbf{S}\mathbf{G} = \left(\begin{array}{c|c} 1 & \boldsymbol{\delta} \\ \hline \boldsymbol{\delta}' & \mathbf{I}_{p_1+p_2} \end{array}\right)$$

unde $\boldsymbol{\delta}=(\delta_1\delta_2)$, $\delta_1=(R_{1.2}, 0,\ldots,0) : p_2\times 1$, $\delta_2=((R^2_{1.23}-R^2_{1.2})^{1/2}, 0,\ldots,0) : p_3\times 1$.

Urmează că

$$Z_2 = R^2_{1.2}, \tag{77}$$

$$Z_3 = (R^2_{1.23} - R^2_{1.2})/(1 - R^2_{1.2}) \tag{78}$$

formează o mulţime de invarianţi maximali.

O mulţime de invarianţi maximali (indusă de grupul de transformări pe spaţiul parametrilor este dată de

$$\gamma_2 = \rho^2_{1.2}, \quad \gamma_3 = (\rho^2_{1.23} - \rho^2_{1.2})/(1 - \rho^2_{1.2}). \tag{79}$$

În termeni de γ_2 şi γ_3 problemele din § 5.1 pot fi scrise ca :

i) $H_1 : \gamma_2 = \gamma_3 = 0$ faţă de $K_1 : \gamma_2 > 0$, $\gamma_3 = 0$
ii) $H_2 : \gamma_2 = \gamma_3 = 0$ faţă de $K_2 : \gamma_2 = 0$, $\gamma_3 > 0$
iii) $H_3 : \gamma_3 = 0$ faţă de $K_3 : \gamma_3 > 0$
iv) $H_4 : \gamma_2 = 0$ faţă de $K_4 : \gamma_2 > 0$

Următoarea lemă dă interpretarea statisticilor Z_3 şi γ_3 :

LEMA 1 (DasGupta (1977)). *Fie $\rho_{13.2}$ coeficientul de corelaţie parţial multiplu (teoretic) dintre X_1 şi X_3 eliminînd efectul regresiei liniare asupra lui X_2. Fie $R_{13.2}$ coeficientul de corelaţie parţial multiplu (de selecţie). Atunci*

$$\gamma_3 = \frac{\rho^2_{1.23} - \rho^2_{1.2}}{1 - \rho^2_{1.2}} = \rho^2_{13.2} \tag{80}$$

$$Z_3 = \frac{R^2_{1.23} - R^2_{1.2}}{1 - R^2_{1.2}} = R^2_{13.2}. \tag{81}$$

Regiunile critice pentru testele raportului de verosimilităţi pentru problemele :

i) şi iv) sînt de forma $Z_2 > k$
ii) şi iii) sînt de forma $Z_3 > k$.

Scanned with OKEN Scanner

.3. Repartiția statisticilor Z_2 și Z_3

Fără a pierde din generalitate presupunem că :

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \mathbf{\Sigma}_{12} & \mathbf{\Sigma}_{13} \\ \mathbf{\Sigma}_{21} & \mathbf{I}_{p_2} & 0 \\ \mathbf{\Sigma}_{21} & 0 & \mathbf{I}_{p_3} \end{pmatrix} \tag{82}$$

Atunci

$$1 - \rho^2_{1.23} = (1 - \gamma_2)(1 - \gamma_3) = 1 - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{21} - \mathbf{\Sigma}_{13}\mathbf{\Sigma}_{31}, \tag{83}$$

$$1 - \rho^2_{1.2} = 1 - \gamma_2 = 1 - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{21}. \tag{84}$$

Fie $g(\cdot\,;\,m)$ densitatea de repartiție a unei variabile $\chi^2(m)$. Definim

$$f_m(y\,;\,s,\,\delta) = (1 - \delta)^{s/2} g(y\,;\,m)\,{}_1F_1(s/2\,;\,m/2\,;\,y\delta/2), \tag{85}$$

pentru $y > 0$, unde $\delta \geqslant 0$ și ${}_1F_1$ este funcția hipergeometrică confluentă

LEMA 2 (DasGupta (1977)). *Fie* $\mathbf{A} : p \times p$ *o matrice aleatoare având o repartiție Wishart* $W_p(m, \mathbf{\Gamma})$ *nesingulară fie* $\mathbf{A}$ *și* $\mathbf{\Gamma}$ *partiționate după cum urmează :*

$$\mathbf{A} = \begin{matrix} \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \begin{matrix} p_1 \\ p_2 \end{matrix} \\ \begin{matrix} p_1 & p_2 \end{matrix} \end{matrix}, \quad \mathbf{\Gamma} = \begin{matrix} \begin{pmatrix} \mathbf{\Gamma}_{11} & \mathbf{\Gamma}_{12} \\ \mathbf{\Gamma}_{21} & \mathbf{\Gamma}_{22} \end{pmatrix} \begin{matrix} p_1 \\ p_2 \end{matrix} \\ \begin{matrix} p_1 & p_2 \end{matrix} \end{matrix}$$

unde $p = p_1 + p_2$ $(m \geqslant p)$. *Atunci*

— $\mathbf{A}_{11.2} \equiv \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$ *și* $(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}, \mathbf{A}_{22})$ *sînt independente*

— $\mathbf{A}_{11.2} \sim W_{p_1}(m - p_2, \mathbf{\Gamma}_{11.2})$, *unde* $\mathbf{\Gamma}_{11.2} = \mathbf{\Gamma}_{11} - \mathbf{\Gamma}_{12}\mathbf{\Gamma}_{22}^{-1}\mathbf{\Gamma}_{21}$.

Dacă $p_1 = 1$, atunci $\mathbf{A}_{11}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}/\mathbf{\Gamma}_{11.2}$ are densitatea de repartiție $f_{p_2}(\cdot\,;\,m,\,\tau^2)$, unde $\tau^2 = \mathbf{\Gamma}_{12}\mathbf{\Gamma}_{22}^{-1}\mathbf{\Gamma}_{21}/\mathbf{\Gamma}_{11}$.

Considerăm $\mathbf{S}$ ca în (82) și definim

$$U_1 = \frac{\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}}{\mathbf{\Sigma}_{11.2}} = \frac{R^2_{1.2}S_{11}}{1 - \rho^2_{1.2}}, \tag{86}$$

unde $\mathbf{\Sigma}_{11.2} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$. Definim

$$\mathbf{S}_{[13].2} \equiv \begin{pmatrix} \mathbf{S}_{11.2} & \mathbf{S}_{13.2} \\ \mathbf{S}_{31.2} & \mathbf{S}_{33.2} \end{pmatrix} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{13} \\ \mathbf{S}_{31} & \mathbf{S}_{33} \end{pmatrix} - \begin{pmatrix} \mathbf{S}_{12} \\ \mathbf{S}_{32} \end{pmatrix} \mathbf{S}_{22}^{-1}(\mathbf{S}_{21}\mathbf{S}_{23}),$$

și analog $\mathbf{\Sigma}_{[13].2}$. Din lema 2 urmează că

i) U_1 și $\mathbf{S}_{[13].2}$ sînt independente

ii) $\mathbf{S}_{[13].2} \sim W_{1+p_2}(n - p_2, \mathbf{\Sigma}_{[13].2})$ cu $n = N - 1$ și

iii) U_1 are densitatea de repartiție $f_{p_2}(\cdot\,;\,n,\,\rho^2_{1.2})$.

Definim

$$U_2 \equiv \frac{\mathbf{S}_{13.2}\mathbf{S}_{33.2}^{-1}\mathbf{S}_{31.2}}{\mathbf{\Sigma}_{11.2} - \mathbf{\Sigma}_{13.2}\mathbf{\Sigma}_{33.2}^{-1}\mathbf{\Sigma}_{31.2}} = \frac{\mathbf{S}_{11}(R^2_{1.23} - R^2_{1.2})}{1 - \rho^2_{1.23}},$$

$$U_3 \equiv \frac{\mathbf{S}_{11.2} - \mathbf{S}_{13.2}\mathbf{S}_{33.2}^{-1}\mathbf{S}_{31.2}}{\mathbf{\Sigma}_{11.2} - \mathbf{\Sigma}_{13.2}\mathbf{\Sigma}_{33.2}^{-1}\mathbf{\Sigma}_{31.2}} = \frac{\mathbf{S}_{11}(1 - R^2_{1.23})}{1 - \rho^2_{1.23}}.$$

Aplicînd lema 2 la $\mathbf{S}_{[13].2}$ obținem

iv) U_2 și U_3 sînt independente

v) $U_3 \sim \chi^2(n - p_2 - p_3)$

vi) U_2 are densitatea de repartiție $f_{p_3}(\cdot\,;\, n - p_2,\ \rho^2_{13.2})$.

Definim

$$W_2 = \frac{U_1}{(1 - \gamma_3)(U_2 + U_3)} = \frac{R^2_{1.2}}{1 - R^2_{1.2}}, \tag{89}$$

$$W_3 = \frac{U_2}{U_2 + U_1} = \frac{R^2_{1.23} - R^2_{1.2}}{1 - R^2_{1.2}}. \tag{90}$$

Atunci densitatea de repartiție comună a statisticilor W_2 și W_3 este:

$$\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\frac{a_j}{j!}\cdot\frac{b_k}{k!}\left\{\frac{(1 - \gamma_3)^j W_2^{p_2/2+j-1}}{B(p_2/2 + j,\ (n - p_2)/2 + k)[1 + (1 - \gamma_3)W_2]^{n/2+j-k}}\right\} \times$$

$$\times \left\{\frac{W_3^{p_3/2+k-1}(1 - W_3)^{(n-p_2-p_3)/2-1}}{B(p_3/2 + k,\ (n - p_2 - p_3)/2)}\right\},$$

unde

$$a_j = \Gamma\left(\frac{n}{2} + j\right)(1 - \gamma_2)^{n/2}\gamma_2^j/\Gamma\left(\frac{n}{2}\right),$$

$$b_k = \Gamma\left(\frac{n - p_2}{2} + k\right)\gamma_3^k(1 - \gamma_3)^{n/2}/\Gamma\left(\frac{n - p_2}{2}\right).$$

Deci densitatea lui $Z_2 = W_2/(1 + W_2)$ și $Z_3 = W_3$ este

$$h(z_2, z_3\,;\, \gamma_2, \gamma_3) \equiv$$

$$\equiv \sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\frac{a_j\ b_k}{j!\ k!}\left\{\frac{(1 - \gamma_3)^j z_2^{p_2/2+j-1}(1 - z_2)^{(n-p_2)/2+k-1}}{B(p_2/2 + j,\ (n - p_2)/2 + k)(1 - \gamma_3 z_2)^{n/2+j+k}}\right\} \times$$

$$\times \left\{\frac{z_3^{p_3/2+k-1}(1 - z_3)^{(n-p_2-p_3)/2-1}}{B(p_3/2 + k,\ (n - p_2 - p_3)/2}\right\}, \quad 0 < z_2,\ z_3 < 1.$$

Scanned with OKEN Scanner

5.4. Teste pentru problemele i)—iv)

Problema i). *Avem*

$$h(z_2, z_3; \gamma_2, 0)/h(z_2, z_3; 0, 0) =$$

$$= \sum_{j=0}^{\infty} \frac{\Gamma\left(\frac{1}{2}n + j\right)}{\Gamma\left(\frac{1}{2}n\right)} \cdot \frac{\gamma_2^j}{j!} \frac{B\left(\frac{1}{2}p_2, \frac{1}{2}(n-p_2)\right)}{B\left(\frac{1}{2}p_2 + j, \frac{1}{2}(n-p_2)\right)} z_2^j.$$

Din lema Neyman-Pearson urmează că testul invariant U.C.M.P. de volum α are regiunea critică

$$z_2 > \lambda(\alpha; p_2, n - p_2), \tag{95}$$

unde $\lambda(\alpha; a, b)$ este punctul superior al repartiției $B\left(\frac{1}{2}a, \frac{1}{2}b\right)$.

Problema ii) *Avem*:

$$h(z_2, z_3; 0, \gamma_3)/h(z_2, z_3; 0, 0) =$$

$$= \sum_{k=0}^{\infty} c_k \gamma_3^k [(1 - z_2)/(1 - \gamma_3 z_2)]^k z_3^k,$$

unde c_k este o constantă numerică pozitivă ce depinde de k și n, p_2, p_3.

Raportul (96), pentru γ_3 și z_2 este o funcție crescătoare de z_3.

Considerăm clasa tuturor testelor condiționate date de

$$\Phi_c(\mathcal{G}) \equiv \{\varphi \in \Phi(\mathcal{G}) : M_{H_2}[\varphi(Z_2, Z_3) | Z_2 = z_2] \leqslant \alpha \text{ aproape pentru toți } z_2\}. \tag{97}$$

Din lema Neyman-Pearson urmează că puterea testului condiționat (dat Z_2) cu regiunea critică

$$z_3 > \lambda(\alpha; p_3, n - p_2 - p_3) \tag{98}$$

este maximă, uniform în K_2, printre toate testele din $\Phi_c(\mathcal{G})$. Deoarece repartiția lui Z_2 este liberă de γ_3, testul anterior este uniform C.M.P. în $\Phi_c(\mathcal{G})$.

Problema iii). Cînd $\gamma_3 = 0$, Z_2 este suficientă și completă. Fie $h^*(z_3, z_2; \gamma_2, \gamma_3)$ densitatea de repartiție a lui Z_3, dat $Z_2 = z_2$ în z_3. Cînd $\gamma_3 = 0$, Z_2 și Z_3 sînt independente. Fie h_3 densitatea de repartiție a lui Z_3 cînd $\gamma_3 = 0$. Atunci.

$$h^*(z_3; z_2, \gamma_2, \gamma_3)/h_3(z) = \sum_{k=0}^{\infty} d_k z_3^k, \tag{99}$$

unde $d_k > 0$ conține z_2, γ_2 și γ_3. Urmează că testul cu regiunea critică (98) este uniform C.M.P. similar la volumul α în $\Phi(\mathcal{G})$. Cînd $p_3 = 1$, testul anterior este uniform C.M.P. nedeplasat de volum α (XIII § 4.).

Pentru a arăta că Z_2 este mărginită complet, notăm că densitatea lui Z_2 poate fi scrisă ca

$$M\left[\frac{z_2^{(1/2)p_2 + J - 1}(1 - z_2)^{(1/2)(n-p_2)-1}}{B\left(\frac{1}{2}p_2 + J, \frac{1}{2}(n-p_2)\right)}\right], \tag{100}$$

Scanned with OKEN Scanner

unde media este luată după J care are repartiția

$$(101) \qquad P(J=j) = \Gamma\left(\frac{1}{2}n+j\right)\gamma_2^j(1-\gamma_2)^{n/2}/\Gamma\left(\frac{1}{2}n\right),$$

$j = 0, 1, \ldots$ Completitudinea lui Z_2 urmează din următoarele fapte
i) Repartiția binominală negativă a lui J este completă.
ii) Șirul de momente a repartiției cu suport mărginit determină unic repartiția

De asemenea completitudinea lui Z_2 urmează din completitudinea repartiției Wishart.

Problema iv) Nu poate fi găsit nici un test optim în $\Phi(\mathcal{G})$ pentru această problemă. Poate prezenta interes stabilirea faptului că testul cu regiunea critică (95) este uniform C.M.P. în $\Phi(\mathcal{G})$ printre toate testele a căror funcție de putere nu depinde de γ_3.

Observație. Rezultatele anterioare pot fi folosite (DasGupta (1969)) pentru a da o regulă optimă în clasa $\mathcal{G}$-invariantă pentru cele 3 probleme de decizie date anterior.

$$H_1 : \gamma_3 > 0, \quad H_2 : \gamma_3 = 0, \quad \gamma_2 > 0, \quad H_3 : \gamma_3 = \gamma_2 = 0.$$

§ 6. Teste în analiza corelației canonice pentru populații care nu sînt normale

ρ_1 din XXVIII §. 13.1.2 este corelația maximă dintre două funcții liniare $y_{11} = \alpha_1'\mathbf{x}_1$ și $y_{12} = \beta_2'\mathbf{x}_2$ supuse la condiția ca $D^2(y_{11}) = D^2(y_{12}) = 1$; atunci ρ_2 este corelația maximă dintre $y_{21} = \alpha_2'\mathbf{x}_1$ și $y_{22} = \beta_2'\mathbf{x}_2$ supuse la condiția că y_{21} și y_{22} sînt necorelate cu y_{11} și y_{12} și au dispersia unu,...

DEFINIȚIA 2. Variabilele y_{i1}, y_{i2}, poartă numele de variabile canonice de ordin i $(i = 1, \ldots, m_1)$.

Cea mai importantă ipoteză în analiza corelației canonice este ipoteza nulă conform căreia cei mai mici $m_1 - k$ coeficienți de corelație canonică sînt zero, adică :

$$H_{m_1-k} : \rho_{k+1} = \ldots = \rho_{m_1} = 0 \qquad (\rho_k > 0).$$

Dacă această ipoteză este adevărată atunci variabilele canonice corespunzătoare y_{i1}, y_{i2} pentru $i = k+1, \ldots, m_1$ nu au valori predictive și relațiile dintre $\mathbf{x}_1$ și $\mathbf{x}_2$ pot fi caracterizate de primele k variabile canonice. În cazul, în care fiecare mulțime conține mai multe variabile, aceasta poate conduce la o reducere substanțială în dimensionalitate.

TEOREMA 5 (Bartlett (1938), (1947)). *Dacă populațiile sînt normale și dacă H_{m_1-k} este adevărată, asimptotic* $(N \to \infty)$,

$$T_{m_1-k} = -N \log l_{m_1-k} \sim \chi^2((m_1-k)(m_2-k))$$

unde

$$l_{m_1-k} = \prod_{i=k+1}^{m_1}(1-r_i^2)$$

este statistica pe care se bazează testul raportului de verosimilități.

Scanned with OKEN Scanner

Pentru a dezvolta o teorie asimptotică privind repartiţia statisticii T_{m_1-k} presupunem că cei mai mari k coeficienţi de corelaţie canonică sînt distincţi.

Scriem statistica test ca

$$T_{m_1-k} = -N\left\{\log\frac{|\mathbf{S}|}{|\mathbf{S}_{11}||\mathbf{S}_{22}|} - \log\prod_{i=1}^{k}(1-r_i^2)\right\}.$$

Folosind dezvoltarea (XXVIII.251) pentru r_i^2, putem scrie

$$T_{m_1-k} = -N\sum_{i=k+1}^{m_1}\log(1-\rho_i^2) + N^{1/2}C + D + \mathbf{0}_p(N^{-1/2}), \tag{102}$$

unde

$$C = \sum_{i=k+1}^{m_1}\rho_i(2z_{i,m_1+i} - \rho_i z_{ii} - \rho_i z_{m_1+i,m_1+i}), \tag{103}$$

$$D = \sum_{i=k+1}^{m_1}\sum_{j=k+1}^{m_2}(z_{i,m_1+j} - \rho_i z_{m_1+i,m_1+j})^2 +$$

$$+\sum_{i=k+1}^{m_1}\sum_{j=k+1}^{m_1}\rho_i\Big\{\rho_j(z_{ij}z_{m_1+i,m_1+j} + z_{i,m_1+j}z_{j,m_1+i}) + \rho_i z_{ij}^2 -$$

$$-\frac{1}{2}\rho_i\rho_j^2(z_{ij}^2 + z_{m_1+i,m_1+j}^2)\Big\} - 2\sum_{i=k+1}^{m_1}\sum_{j=1}^{k}\rho_i\delta_{ij}z_{i,m_1+j}(\rho_j z_{ij} + \rho_i z_{m_1+i,m_1+j}) -$$

$$-\sum_{i=1}^{k}\sum_{j=k+1}^{m_1}\delta_{ji}\{\rho_j^3(z_{ij}^2 + z_{m_1+i,m_1+j}^2) + z_{im_1+j}(2\rho_i z_{j,m_1+i} - 2\rho_i\rho_j z_{m_1+i,m_1+j} +$$

$$+\rho_j z_{i,m_1+j} - 2\rho_j^2 z_{ij}) + 2\rho_i\rho_j^2 z_{ij}z_{m_1+i,m_1+j}\},$$

cu $\delta_{ij} = \rho_i(1-\rho_j^2)/(\rho_i^2 - \rho_j^2)$ şi z_{ij} definite în (XXVIII.256).

Observaţii i). Dezvoltarea (102) este adevărată independent dacă ipoteza H_{m_1-k} este adevărată sau nu, cu condiţia ca coeficienţii de corelaţie canonică nenuli sînt distincţi; este de asemenea adevărată cînd $k = 0$.

Cînd este adevărată ipoteza H_{m_1-k}, (102) se reduce la

$$T_{m_1-k} = \sum_{i=k+1}^{m_1}\sum_{j=k+1}^{m_2} z_{i,m_1+j}^2 + \mathbf{0}_p(N^{-1/2}), \tag{104}$$

unde aici

$$z_{i_1 m_1+j} = \sqrt{N}\,s_{i,m_1+j}.$$

ii). Suma din (104) conţine numai $(m_1-k)(m_2-k)$ elemente ale submatricei $\mathbf{S}_{12}$ ce corespund la coeficienţii de corelaţie ai populaţiei ce sînt zero.

Scanned with OKEN Scanner

Repartiția asimptotică nulă a statisticii T_{m_1-k} se obține din (104) și din repartiția limită a variabilelor z_{i,m_1+j}. Definim vectorul $\mathbf{z}: (m_1-k)(m_2-k)\times 1$ prin

$$\mathbf{z} = (z_{k+1,m_1+k+1}, \ldots, z_{k+1,m_1+m_2}, z_{k+2,m_1+k+1}, \ldots, z_{k+2,m_1+m_2}, \ldots, z_{m_1,m_1+k+1}, \ldots, z_{m_1,m_1+m_2})'. \tag{105}$$

Din teorema limită centrală multidimensională, repartiția limită a vectorului $\mathbf{z}$ (dacă este adevărată ipoteza H_{m_1-k}) este normală cu vectorul valoare medie $\mathbf{0}$ și matricea de covarianțe Ω specificată de

$$D^2(z_{i,m_1+j}) = 1 + \varkappa_{22}^{i,m_1+j}, \quad \operatorname{cov}(z_{i,m_1+j}, z_{i,m_1+l}) = \varkappa_{211}^{i,m_1+j,m_1+l} \quad (j \neq l)$$

$$\operatorname{cov}(z_{i,m_1+j}, z_{l,m_1+j}) = \varkappa_{211}^{m_1+j,il} \ (i \neq l), \quad \operatorname{cov}(z_{i,m_1+j}, z_{l,m_1+p}) = \tag{106}$$

$$= \varkappa_{111}^{i,m_1+j,l,m_1+p} \quad (i \neq l, \ j \neq p).$$

Punînd $\mathbf{v} = \mathbf{Hz}$, unde $\mathbf{H}$ este matricea ortogonală cu

$$\mathbf{H}\Omega\mathbf{H}' = \operatorname{diag}(\omega_1, \ldots, \omega_{(m_1-k)(m_2-k)}) \tag{107}$$

atunci

$$T_{m_1-k} = \sum_{i=1}^{(m_1-k)(m_2-k)} v_i^2 + \mathbf{0}_p(N^{-1/2}),$$

unde elementele v_i ale lui $\mathbf{v}$ sînt asimptotic independente și repartizate $N(0, \omega_i)$, ω_i dat de (107). În concluzie are loc.

TEOREMA 6 (Muirhead, Waternaux (1980)). *Fie r_i coeficienții de corelație canonică ai unei selecții de volum N asupra unei populații multidimensionale avînd cumulanții de ordinul 4 finiți și fie T_{m_1-k} definită de*

$$T_{m_1-k} = -N \sum_{i=k+1}^{m_1} \log(1 - r_i^2).$$

Dacă este adevărată ipoteza H_{m_1-k}, repartiția limită ($N\to\infty$) a statisticii T_{m_1-k} este aceeași cu a statisticii

$$\sum_{i=1}^{(m_1-k)(m_2-k)} \omega_i X_i,$$

unde X_i sînt independente și repartizate $\chi^2(1)$ și ω_i sînt rădăcinile caracteristice ale matricei Ω specificate de (106).

COROLARUL 1. *Pentru repartiția eliptică cu excesul $3\varkappa$, T_{m_1-k} are asimptotic, aceeași repartiție ca o variabilă $\chi^2(m_1-k)(m_2-k)$.*

COROLARUL 2. *Dacă vectorii $\mathbf{x}$ și*

$$\mathbf{x}_\varepsilon = (\varepsilon_1 x^{(1)}, \ldots, \varepsilon_{m_1+m_2} x^{(m_1+m_2)})' \quad (\varepsilon_i = \pm 1)$$

Scanned with OKEN Scanner

au aceeași repartiție (Modelul II) atunci repartiția limită a statisticii T_{m_1-k} *este aceeași cu a unei sume ponderate de* $(m_1 - k)(m_2 - k)$ *variabile aleatoare independente repartizate* $\chi^2(1)$ *cu ponderile*

$$1 + \mathfrak{K}_{22}^{i,m_1+j} \quad (i = k+1, \ldots, m_1;\ j = k+1, \ldots, m_2).$$

Dacă repartițiile bidimensionale sînt identice atunci

$$\frac{T_{m_1-k}}{1 + \mathfrak{K}_{22}} \xrightarrow[N\to\infty]{R} \chi^2((m_1 - k)(m_2 - k)).$$

Observație. Acum Ω este diagonală cu elementele egale cu ponderile.

COROLARUL 3. *Dacă componentele lui* $\mathbf{x}$ *sînt independente (Modelul III),* $\Omega = \mathbf{I}$ *și*

$$T_{m_1-k} \xrightarrow[N\to\infty]{R} \chi^2(m_1 - k)(m_2 - k).$$

Observație. Din corolarele 1 și 2 urmează că dacă $\hat{\mathfrak{K}}$ și $\hat{\mathfrak{K}}_{22}$ sînt estimații consistente pentru $\mathfrak{K}$ și respectiv $\mathfrak{K}_{22}$ atunci repartițiile limită nu se schimbă dacă aceste estimații sînt folosite în locul valorilor parametrilor.

Repartiția statisticii T_{m_1-k} în ipoteza nenulă este dată în :

TEOREMA 7 (Muirhead, Waternaux (1980)). *Dacă populația este eliptică cu excesul* $3\mathfrak{K}$, *atunci :*

a) *În alternativa* $\rho_{k+1} \geqslant \ldots \geqslant \rho_{m_1} \geqslant 0$, $\rho_{k+1} > 0$, *repartiția limită a statisticii*

$$V_{m_1-k} = N^{1/2}\left\{T_{m_1-k} + N\sum_{i=k+1}^{m_1} \log(1 - \rho_i^2)\right\}$$

este normală cu valoarea medie zero și dispersia

$$4(\mathfrak{K} + 1)\sum_{i=k+1}^{m_1} \rho_i^2.$$

b) *Sub șirul de alternative* $K_{m_1-k,N}$: $\rho_i^2 = \theta_i^2/N$ $(i = k+1, \ldots, m_1)$, *repartiția limită a statisticii* $T_{m_1-k}/(1 + \mathfrak{K})$ *este* χ^2 *necentrată cu* $(m_1 - k)(m_2 - k)$ *grade de libertate și parametru de necentricitate*

$$\sum_{i=k+1}^{m_1} \theta_i^2/(1 + \mathfrak{K}).$$

Cînd $\mathfrak{K} = 0$, a) și b) se reduc la rezultatele date de Fujikoshi (1976) pentru selecții normale.

Muirhead, Waternaux (1980) fac un studiu Monte Carlo pentru a determina reparațiile statisticilor T_{m_1-k} pentru populații eliptice : normală, normală contaminată ($\varepsilon = 0{,}1$; $\sigma = 3$) și T pentru 5 grade de libertate. Obțin repartițiile statisticilor $T_{m_1-k}(\mathfrak{K} + 1)$, unde $3\mathfrak{K}$ măsoară excesul populației și $T^*_{m_1-k} = T_{m_1-k}/(\hat{\mathfrak{K}} + 1)$, unde $\hat{\mathfrak{K}}$ este un estimator consistent pentru $\mathfrak{K}$, cînd este adevărată ipoteza.

Scanned with OKEN Scanner

În fiecare caz au fost generate 200 selecţii de volum $N = 100$ de un vector $\mathbf{x} : 7 \times 1$ ($m_1 = 3$, $m_2 = 4$) şi s-a făcut analiza canonică. Pentru fiecare repartiţie au fost considerate două cazuri :

(a) $\boldsymbol{\Sigma} = \mathbf{I}_7$, adică $k = 0$

(b) $\boldsymbol{\Sigma}_{11} = \mathbf{I}_3$, $\boldsymbol{\Sigma}_{22} = \mathbf{I}_4$, $\mathbf{P} = [\mathrm{diag}(0, 8, 0, 0) : 0]$ adică $k = 1$.

În ipoteza $H_{m_1-k} : \rho_{k+1} = \ldots = \rho_{m_1} = 0$, repartiţia comună a celor $r = 2(m_1 - k)$ variabile canonice y_{i1}, y_{i2} ($i = k+1, \ldots, m_1$) este sferică.

Fie

$$R^2 = \sum_{i=k+1}^{m_1} (y_{i1}^2 + y_{i2}^2)$$

raza repartiţiei : atunci ca pentru repartiţia radială a oricărei populaţii sferice r-dimensionale cu dispersia λ şi excesul $3\varkappa$,

$$M(R^2) = r\lambda, \quad M(R^4) = r(r+2)(\varkappa+1)\lambda^2.$$

Astfel $\varkappa$ satisface

$$(\varkappa + 1) = \frac{r}{r+2} \cdot \frac{M(R^4)}{\{M(R^2)\}^2},$$

şi estimatorul consistent este

$$\hat{\varkappa} = \frac{r}{r+2} \cdot \frac{m_4}{m_2^2} - 1,$$

unde

$$m_4 = N^{-1} \sum_{\alpha=1}^{N} \left(\sum_{i=k+1}^{m_1} y_{i1\alpha}^2 + y_{i2\alpha}^2 \right)^2, \quad m_2 = N^{-1} \sum_{\alpha=1}^{N} \sum_{i=k+1}^{m_1} (y_{i1\alpha}^2 + y_{i2\alpha}^2),$$

şi $y_{i1\alpha}$ şi $y_{i2\alpha}$ reprezintă scorurile canonice de ordin i pentru observaţia α.

§ 7. Teste pentru ipoteze asupra matricilor de covarianţe pentru populaţii care nu sînt normale

7.1. Test pentru verificarea ipotezei conform căreia matricea de covarianţe este egală cu o matrice dată

Dată o selecţie de volum N asupra unui vector m-dimensional cu matricea de covarianţe $\boldsymbol{\Sigma}$, ne propunem să considerăm ipoteza nulă

$$H_{01} : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$$

unde $\boldsymbol{\Sigma}_0$ este o matrice simetrică pozitiv definită dată.

Ipoteza H_{01} este echivalentă cu ipoteza nulă

$$H_{01}^* : \boldsymbol{\Sigma} = \mathbf{I}_m .$$

Scanned with OKEN Scanner

Dacă vectorul $\mathbf{x} = (x_1, \ldots, x_N)$ este repartizat normal, statistica pe care se bazează testul raportului de verosimilitate este

$$l_1 = e^{mN/2} |\mathbf{S}|^{N/2} \exp\left\{-\frac{1}{2}\mathrm{Ur}(\mathbf{NS})\right\} \quad \text{(propoziția XVIII 2.)}$$

unde

$$\mathbf{S} = N^{-1} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})'. \tag{108}$$

Dacă scriem

$$\mathbf{S} = \mathbf{I}_m + N^{-1/2}\mathbf{Z},$$

statistica $-2\log l_1$, cînd H_{01}^* este adevărată, are următoarea dezvoltare

$$-2\log l_1 = \frac{1}{2}\mathrm{Ur}(\mathbf{Z}) + \mathbf{0}_p(N^{-1/2}) = \mathbf{u}'\mathbf{u} + \mathbf{0}_p(N^{-1/2}),$$

unde

$$\mathbf{u}' = (z_{11}/2^{1/2}, \ldots, z_{mm}/2^{1/2}, z_{12}, \ldots, z_{1m}, z_{23}, \ldots, z_{m-1,m}).$$

Dacă presupunem că $x_1, \ldots, x_N$ este o selecție dintr-o populație eliptică cu excesul $3\varkappa$, atunci $\mathbf{u}$ are la limită ($N \to \infty$) o repartiție normală $\frac{1}{2}m(m+1)$ dimensională cu vectorul valoare medie zero și matricea de covarianțe:

$$\mathbf{V}_1 = \begin{pmatrix} (1+\varkappa)\mathbf{I}_m + \frac{1}{2}\varkappa\mathbf{1}\mathbf{1}' & \mathbf{0} \\ \mathbf{0} & (1+\varkappa)\mathbf{I}_{\frac{1}{2}m(m-1)} \end{pmatrix}$$

unde $\mathbf{1}' = (1, \ldots, 1) \in R^m$. Deoarece rădăcinile caracteristicice ale matricei $\mathbf{V}_1$ sînt $1 + \varkappa + \frac{1}{2}m\varkappa$ și $1+\varkappa$ repetată de $\frac{1}{2}(m-1)(m-2)$ urmază că dacă este adevărată ipoteza H_{01}^*, $-2\log l_1/(1+\varkappa)$ este repartizată, la limită, ca variabila aleatoare.

$$\left(1 + \frac{1}{2}\frac{\varkappa m}{1+\varkappa}\right) X_1 + X_2$$

unde $X_1 \sim \chi^2(1)$, $X_2 \sim \chi^2\left(\frac{1}{2}(m-1)(m-2)\right)$ și X_1 și X_2 sînt independente (Muirhead, Waternaux) (1980)).

Pentru populații normale $\varkappa = 0$ și asimptotic $-2\log l_1 \sim \chi^2\left(\frac{1}{2}m(m+1)\right)$ (teorema XVIII.12).

Scanned with OKEN Scanner

7.2. Teste pentru sfericitate

Dacă vectorul $\mathbf{x}$ este repartizat normal, statistica pe care se bazează testul raportului de verosimilitate pentru verificarea ipotezei

$H_{02}: \boldsymbol{\Sigma} = \lambda \mathbf{I}_m$, unde λ este necunoscut, este

$$l_2 = |\mathbf{S}|^{N/2} \Big/ \left\{ \frac{1}{m} \operatorname{Ur}(\mathbf{S}) \right\}^{mN/2} \quad \text{(teorema XVIII.18).}$$

Cu $\mathbf{S} = \lambda(\mathbf{I}_m + N^{-1/2}\mathbf{Z})$, cînd H_{02} este adevărată, exprimăm $-2\log l_2$ ca (Muirhead, Waternaux (1980)).

$$-2\log l_2 = \frac{1}{2} \operatorname{Ur}(\mathbf{Z}^2) - (2m)^{-1}\{\operatorname{Ur}(\mathbf{Z})\}^2 + \mathbf{0}_p(N^{-1/2}) =$$

$$= \sum_{i<j} z_{ij}^2 + \frac{1}{2} \sum_{i=1}^{m} (z_{ij} - \bar{z})^2 + \mathbf{0}_p(N^{-1/2}),$$

unde $\bar{z} = m^{-1} \sum z_{ii}$. Dacă selecția provine dintr-o populație eliptică, repartiția limită ($N \to \infty$) a statisticii $-2\log l_2/(1 + \varkappa)$ este $\chi^2\left(\frac{1}{2}(m-1)(m+2)\right)$.

7.3. Test pentru egalitatea a două matrici de covariante

Pentru selecții de volume N_1 și N_2 asupra unor vectori normali m-dimensionali cu matricile de covariante $\boldsymbol{\Sigma}_1$ și $\boldsymbol{\Sigma}_2$, raportul de verosimilități pentru verificarea ipotezei $H_{03}: \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ este

$$l_3 = |\mathbf{S}|^{N_1/2} |\mathbf{S}_2|^{N_2/2} / |\mathbf{S}|^{N/2},$$

unde $\mathbf{S}_1$ și $\mathbf{S}_2$ sînt două matrici de covariante de selecție definite ca în (108) și

$$\mathbf{S} = N^{-1}(N_1 S_1 + N_2 S_2) \text{ cu } N = N_1 + N_2.$$

Fie $N_i = k_i N$ $(i = 1, 2)$ cu $k_1 + k_2 = 1$ și scriem

$$\mathbf{S}_i = \boldsymbol{\Sigma}_0 + (Nk_i)^{-1/2}\mathbf{Z}_i$$

unde $\boldsymbol{\Sigma}_0$ este valoarea comună a lui $\boldsymbol{\Sigma}_1$ și $\boldsymbol{\Sigma}_2$. Dacă este adevărată ipoteza H_{03}, atunci statistica $-2\log l_3$ admite următoarea dezvoltare

$$-2\log l_3 = \frac{1}{2} k_2 \operatorname{Ur}(\mathbf{Z}_1\boldsymbol{\Sigma}_0^{-1})^2 + \frac{1}{2} k_1 \operatorname{Ur}(\mathbf{Z}_2\boldsymbol{\Sigma}_0^{-1})^2 -$$

$$- (k_1 k_2)^{1/2} \operatorname{Ur}(\mathbf{Z}_1\boldsymbol{\Sigma}_0^{-1}\mathbf{Z}_2\boldsymbol{\Sigma}_0^{-1}) + \mathbf{0}_p(N^{-1/2}).$$

Scanned with OKEN Scanner

Acum presupunem că cele două selecţii provin din populaţii eliptice cu excesul $3\varkappa$. Atunci putem scrie

$$-2\log l_3 = \mathbf{v}'\mathbf{v} + \mathbf{0}_p(N^{-1/2}),$$

unde

$$\mathbf{v} = (1+\varkappa)^{-1/2}(\boldsymbol{\Sigma}_0 \otimes \boldsymbol{\Sigma}_0)^{-1/2}\left\{\left(\frac{1}{2}k_2\right)^{1/2}\mathbf{z}_1 - \left(\frac{1}{2}k_1\right)^{1/2}\mathbf{z}_2\right\}$$

unde $\mathbf{z}_1$ şi $\mathbf{z}_2$ sînt vectori $m^2 \times 1$ formaţi din $\mathbf{Z}_1$ şi respectiv $\mathbf{Z}_2$ prin operaţii asupra liniilor lor. Repartiţia asimptotică a statisticii $\mathbf{v}$ $(N \to \infty)$ este normală m^2-dimensională cu vectorul valoarea medie $\mathbf{0}$ şi matricea din covarianţă $\mathbf{V}_2$ dată de

$$\mathbf{V}_2 = \frac{\varkappa}{2(1+\varkappa)}\sum_{i,j=1}^{m}(E_{ij} \otimes E_{ij}) + \mathbf{P}$$

unde

$$\mathbf{P} = \frac{1}{2}\left\{\mathbf{I}_{m^2} + \sum_{i,j=1}^{m}(\mathbf{E}_{ij} \otimes \mathbf{E}'_{ij})\right\},$$

$\mathbf{E}_{ij}$ fiind o matrice $m \times m$ cu unu în poziţia (i, j) şi zero în rest. Rangul lui $\mathbf{V}_2$ este $\frac{1}{2}m(m+1)$. Urmează că în cazul în care este adevărată ipoteza H_{03}, $-2\log l_3/(1+\varkappa)$ are asimptotic aceeaşi repartiţie ca o sumă ponderată de $\frac{1}{2}m(m+1)$ variabile aleatoare independente repartizate $\chi^2(1)$ ponderile fiind rădăcinile caracteristice ale lui $\mathbf{V}_2$. Cînd cele două selecţii provin din populaţii normale cu $\varkappa = 0$ şi asimptotic matricea de covarianţe $\mathbf{V}_2$ este egală cu $\mathbf{P}$ care este indempotentă, obţinem că la limită $-2\log l_3 \sim \chi^2\left(\frac{1}{2}m(m+1)\right)$. Nu este posibil ca în cazul mai general al populaţiilor eliptice să ajustăm statistica test astfel încît repartiţia limită să concorde cu cea obţinută în ipoteza de normalitate.

Rezultate similare se obţin cînd verificăm egalitatea mai multor matrici de covarianţe.

7.4. Test pentru verificarea ipotezei conform căreia $\boldsymbol{\Sigma}$ este o matrice diagonală bloc

Pentru o selecţie de ordin N asupra unui vector normal m-dimensional raportul de verosimilităţi pentru verificarea ipotezei

$$H_{04} : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 = \operatorname{diag}(\boldsymbol{\Sigma}_{11}, \ldots, \boldsymbol{\Sigma}_{qq})$$

unde $\boldsymbol{\Sigma}_{ii}$ este $m_i \times m_i$ $(i = 1, \ldots, q)$ cu $\sum m_i = m$, este

$$l_4 = |\mathbf{S}|^{N/2}\Big/\prod_{i=1}^{q}|\mathbf{S}_{ii}|^{N/2},$$

Scanned with OKEN Scanner

unde $\mathbf{S}$ este definită de (108) și partiționată în submatrici $\mathbf{S}_{ij} : m_i \times m_i$. Scriind $\mathbf{S} = \boldsymbol{\Sigma}_0 + N^{-1/2}\mathbf{Z}$, cînd H_{04} este adevărată, $-2\log l_4$ admite dezvoltarea

$$-2\log l_4 = \sum_{i<j}^{q} \mathrm{Ur}(\mathbf{Z}_{ij}\boldsymbol{\Sigma}_{jj}^{-1}\mathbf{Z}_{ji}\boldsymbol{\Sigma}_{ii}^{-1}) + 0_p(N^{-1/2}) = \sum_{i<j}^{q} \mathbf{z}_{ij}'(\boldsymbol{\Sigma}_{ii} \otimes \boldsymbol{\Sigma}_{jj})^{-1}\mathbf{z}_{ij} + \mathbf{0}_p(N^{-1/2}),$$

unde $\mathbf{z}_{ij}$ este un vector $m_i m_j \times 1$, obținut transformînd liniile lui $\mathbf{Z}_{ij}$. Dacă presupunem că observațiile provin dintr-o populație eliptică cu excesul $3\varkappa$ atunci $\mathbf{z}_{ij}$ $(i < j)$ sînt asimptotic independente și repartiția limită (cînd $N \to \infty$) a variabilelor $\mathbf{z}_{ij}$ este normală $m_i m_j$-dimensională cu vectorul valoare medie $\mathbf{0}$ și matricea de covariante $(1 + \varkappa)(\boldsymbol{\Sigma}_{ii} \otimes \boldsymbol{\Sigma}_{jj})$.

Dacă este adevărată ipoteza H_{04}, atunci (Muirhead, Waternaux (1980), $-2\log l_4/(1 + \varkappa)$ are la limită, o repartiție $\chi^2(f)$, unde $f = \sum_{i<j} m_i m_j$. Cînd $q = 0$, se obține rezultatul corolarului 1 al teoremei 6.

Scanned with OKEN Scanner

Capitolul XXX

VALORI MEDII CONDIȚIONATE. REGRESIE

§ 1. Curbe de regresie

Fie $f(x, y)$ densitatea de repartiție comună a variabilelor X și Y.

DEFINIȚIA 1. Valoarea medie a variabilei Y condiționată de $X = x$, definită de

$$\mu'_{1x} = M(Y \mid x) = \int_{-\infty}^{\infty} yf(x, y)\,dy \Big/ \int_{-\infty}^{\infty} f(x, y)\,dy \tag{1}$$

poartă numele de ecuația curbei de regresie a variabilei Y asupra lui X.

Valoarea medie a variabilei X condiționată de $Y = y$, definită de

$$\mu'_{1y} = M(X \mid y) = \int_{-\infty}^{\infty} xf(x, y)\,dx \Big/ \int_{-\infty}^{\infty} f(x, y)\,dx \tag{2}$$

poartă numele de ecuația curbei de regresie a variabilei X asupra lui Y.

DEFINIȚIA 2. μ'_{rx} definită de

$$\mu'_{rx} = M(Y^r \mid x) = \int_{-\infty}^{\infty} y^r f(x, y)\,dy \Big/ \int_{-\infty}^{\infty} f(x, y)\,dy \tag{3}$$

poartă numele de ecuația curbei de regresie de ordinul r a variabilei Y asupra lui X

$$\mu'_{ry} = M(X^r \mid y) = \int_{-\infty}^{\infty} x^r f(x, y)\,dy \Big/ \int_{-\infty}^{\infty} f(x, y)\,dx \tag{4}$$

poartă numele de ecuația curbei de regresie de ordinul r a variabilei X asupra lui Y.

DEFINIȚIA 3. Momentul

$$\mu_{rx} = M[\{Y - M(Y \mid x)\}^r \mid x] =$$

$$= \int_{-\infty}^{\infty} \{y - M(Y \mid x)\}^r f(x, y)\,dy \Big/ \int_{-\infty}^{\infty} f(x, y)\,dy \tag{5}$$

Scanned with OKEN Scanner

caracterizează dependența momentului central al variabilei Y de x

(5) cu $r = 2$ poartă numele de ecuația curbei scedastică

(5) cu $r = 3$ poartă numele de ecuația curbei clitică

(5) cu $r = 4$ poartă numele de ecuația curbei kurtică.

§ 2. Regresie liniară. Coeficienți de regresie

DEFINIȚIA 4. Dacă X și Y sînt două variabile aleatoare, momentul produs sau covarianța dintre X și Y este definită de

$$\mu_{11} = \operatorname{cov}(X, Y) = M\{[X - M(X)][Y - M(y)]\} = \tag{6}$$
$$= M(XY) - M(X)M(Y) = M(XY) - \mu_1\mu_2.$$

PROPOZIȚIA 1. *Dacă X și Y sînt variabile aleatoare independente atunci*

$$\mu_{11} = \mathfrak{M}_{11} = 0. \tag{7}$$

Exemplul 1. Dacă (X, Y) au o repartiție normală bidimensională, funcția caracteristică comună a variabilelor X^2 și Y^2 este

$$\varphi(t, u) = \frac{1}{2\pi(1 - \rho^2)^{1/2}} \iint \exp\left\{-\frac{1}{2(1 - \rho^2)}[x^2\{1 - 2(1 - \rho^2)it\} - 2\rho xy + \right.$$
$$\left. + y^2\{1 - 2(1 - \rho^2)iu\}]\right\} dx\, dy =$$
$$= (1 - \rho^2)^{1/2}[\{1 - 2(1 - \rho^2)it\}\{1 - 2(1 - \rho^2)iu\} - \rho^2]^{-1/2} =$$
$$= [(1 - 2it)(1 - 2iu) + 4\rho^2 tu]^{-1/2}.$$

Avem
$$\varphi(t, 0) = (1 - 2it)^{-1/2}, \quad \varphi(0, u) = (1 - 2iu)^{-1/2}$$

ceea ce arată că repartițiile marginale sînt $\chi^2(1)$.

De asemenea
$$\mu_{11} = \mathfrak{M}_{11} = \left[\frac{\partial^2 \log \varphi(t, u)}{\partial(it)\,\partial(iu)}\right]_{t=u=0} = 2\rho^2.$$

Cînd $\rho = 0$,
$$\varphi(t, u) = \varphi(t, 0)\varphi(0, u)$$

ceea ce reprezintă condiția necesară și suficientă pentru independența lui X^2 și Y^2. Astfel în acest caz independența implică $\mu_{11} = 0$.

Observație. În general, reciproca nu este adevărată.

Exemplul 2. Considerăm repartiția uniformă bidimensională pe cercul unitate centrată în $M(X)$ și $M(Y)$. Avem.

$$dF(x, y) = dx\, dy/\pi, \quad 0 \leqslant x^2 + y^2 \leqslant 1,$$

Scanned with OKEN Scanner

de unde

$$\mu_{11} = \iint xy\,dF = \frac{1}{\pi}\iint xy\,dx\,dy =$$

$$= \frac{1}{\pi}\int x\left[\int y\,dy\right]dx =$$

$$= \frac{1}{\pi}\int x\left[\frac{1}{2}y^2\right]_{-(1-x^2)^{1/2}}^{(1-x^2)^{1/2}} dx = 0.$$

Însă X și Y nu sînt independente, deoarece domeniul de variație al fiecăreia depinde de valorile celeilalte.

DEFINIȚIA 5. Vom spune că regresia lui $X(Y)$ asupra lui $Y(X)$ este liniară dacă

$$M(X \mid y) = \alpha_1 + \beta_1 y \left(M(Y \mid x) = \alpha_2 + \beta_2 x\right) \tag{8}$$

unde $\alpha_1(\alpha_2)$ și $\beta_1(\beta_2)$ sînt constante.

PROPOZIȚIA 2. *Avem*

$$\beta_1 = \mu_{11}/\sigma_2^2, \quad \beta_2 = \mu_{22}/\sigma_1^2. \tag{9}$$

Demonstrație. Arătăm că $\beta_2 = \mu_{22}/\sigma_1^2$. Dacă notăm prin $f(y \mid x)(f_Z(z))$ densitatea de repartiție a variabilei Y condiționate de $X = x$ (densitatea de repartiție a variabilei Z) avem

$$\int M(Y \mid x) f_X(x)\,dx = \alpha_2 + \beta_2 \int x f_X(x)\,dx$$

sau

$$M(Y) = \alpha_2 + \beta_2 M(X). \tag{10}$$

De asemenea

$$\int M(Y \mid x) x f_X(x)\,dx = \alpha_2 \int x f_X(x)\,dx + \beta_2 \int x^2 f_X(x)\,dx$$

sau

$$M(XY) = \alpha_2 M(X) + \beta_2 M(X^2). \tag{11}$$

Rezolvînd sistemul format din ecuațiile (10) și (11) în raport cu α_2 și β_2, obținem:

$$\beta_2 = \frac{M(XY) - M(X)M(Y)}{M(X^2) - M^2(X)} = \frac{\mu_{11}}{\sigma_1^2}, \quad \alpha_2 = M(Y) - \beta_2 M(X).$$

Similar găsim pe β_1,

DEFINIȚIA 6. $\beta_1(\beta_2)$ poartă numele de coeficientul de regresie a lui $X(Y)$ asupra lui $Y(X)$.

Scanned with OKEN Scanner

Din (8), (12) găsim

$$M(Y \mid x) - \mu_2 = \beta_2(x - \mu_1) \tag{13}$$

şi similar

$$M(X \mid y) - \mu_1 = \beta_1(y - \mu_2). \tag{14}$$

DEFINIŢIA 7. (13) şi (14) poartă numele de ecuaţiile regresiei liniare.

Exemplul 3. Regresia lui $X^2(Y^2)$ asupra lui $Y^2(X^2)$ este liniară (*exemplul 1*).

Din (XXVIII.8) cu $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = \sigma_2^2 = 1$,

$$M(X \mid y) = \rho y, \quad D^2(X \mid y) = 1 - \rho^2.$$

Astfel

$$M(X^2 \mid y) = D^2(X \mid y) + \{M(X \mid y)\}^2 = 1 - \rho^2 + \rho^2 y^2.$$

La fiecare valoare a lui y^2 corespund valorile $+y$ şi $-y$ ce apar cu probabilităţi egale. Astfel, deoarece $M(X^2 \mid y)$ este funcţie numai de y

$$M(X^2 \mid y^2) = \frac{1}{2}\{M(X^2 \mid y) + M(X^2 \mid -y)\} = M(X^2 \mid y) = 1 - \rho^2 + \rho^2 y^2,$$

expresie pe care s-o putem rescrie în forma (14)

$$M(X^2 \mid y^2) - 1 = \rho^2(y^2 - 1)$$

şi regresia lui X^2 asupra Y^2 este strict liniară. La fel,

$$M(Y^2 \mid x^2) - 1 = \rho^2(x^2 - 1).$$

Cum $\mu_{11} = 2\rho^2$ (exemplul 1) şi dispersiile sînt egale cu 2 (deoarece X^2 şi Y^2 sînt repartizate $\chi^2(1)$) urmează (din (9)) că ρ^2 este coeficientul de regresie în fiecare din ecuaţiile regresiei liniare.

Exemplul 4. În exemplul 2, $M(X \mid y) = M(Y \mid x) = 0$, astfel că avem regresii liniare, coeficienţii fiind zero.

Exemplul 5. Considerăm variabilele X şi Y^2 din exemplul 3. Am văzut că regresia lui Y^2 asupra lui X^2 este liniară cu coeficientul ρ^2 şi prin urmare cînd $\rho \neq 0$ nu este liniară asupra lui X. Totuşi, deoarece $M(X \mid y) = \rho y$, avem

$$M(X \mid y^2) = \frac{1}{2}\{M(X \mid y) + M(X \mid -y)\} = 0,$$

astfel că regresia lui X asupra lui Y^2 este liniară cu coeficientul de regresie 0.

§ 3. Coeficientul de corelaţie

PROPOZIŢIA 3. *Coeficientul de corelaţie dintre variabilele aleatoare X şi Y notat prin ρ şi definit ca*

$$\rho = \mu_{11}/\sigma_1\sigma_2$$

Scanned with OKEN Scanner

are următoarele proprietăți

i) *este o funcție simetrică de* X *și* Y

ii) *este invariant la translații și omotetii.*

iii) *are același semn ca* β_1 *și* β_2

iv) $|\rho|$ *este media geometrică a* $|\beta_1|$ *și* $|\beta_2|$

v) $0 \leqslant \rho^2 \leqslant 1$, ρ^2 *fiind egal cu 1 dacă și numai dacă* $(x - \mu_1)$ *și* $(y - \mu_2)$ *sînt proporționale strict*

vi) *Dacă* X *și* Y *sînt independente,* $\rho = 0$.

Demonstrație. ii) rezultă din faptul că ρ este o funcție omogenă de momentele față de medii iii) ρ, β_1 și β_2 au ca numărător pe μ_{11} și numitori pozitivi. iv) Din (9) și din definiția coeficientului de corelație, avem

$$\rho^2 = \beta_1 \beta_2. \tag{15}$$

v) Din inegalitatea lui Schwarz

$$\mu_{11}^2 = \left\{\iint (x - \mu_1)(y - \mu_2)\, dF\right\}^2 \leqslant \left\{\int (x - \mu_1)^2\, dF\right\}\left\{\int (y - \mu_2)^2\, dF\right\} = \sigma_1^2 \sigma_2^2,$$

ceea ce demonstrează afirmația făcută vi) rezultată din proporziția 3.

PROPOZIȚIA 4. *Unghiul dintre dreptele de regresie* (13) *și* (14) *este dat de*

$$\theta = \operatorname{arctg}\left(\frac{1 - \rho^2}{\rho} \frac{\sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2}\right). \tag{16}$$

Observație. Cînd ρ variază de la -1 la $+1$, θ crește de la 0 pînă la maximul său de $\frac{1}{2}\pi$ cînd $\rho = 0$ și apoi descrește din nou la 0. Astfel, dacă și numai dacă X și Y sînt într-o relație funcțională strict liniară, cele două drepte de regresie coincid ($\rho^2 = 1$). Dacă și numai dacă $\rho = 0$, caz în care X și Y poartă numele de variabile necorelate, cele două drepte de regresie fac unghiuri drepte.

Observație. Din faptul că reciproca proprietății vi) din propoziția 3 nu este în general adevărată, rezultă diricultatea interpretării lui ρ ca coeficientul de interdependență. De fapt ρ este, coeficient al interdependenței liniare și formele mai complicate de interdependență variază în afara acestei terminologii. În general problema variației comune este atît de complexă încît nu poate fi înțeleasă cu un singur coeficient.

A *exprima* o calitate nu este același lucru cu a *o măsura*. Rămîne o problemă deschisă ce funcție de ρ va fi folosită ca *măsură* a independenței. Cum

$$\rho^2 = D^2(\alpha_2 + \beta_2 x)/\sigma_2^2 = D^2(\alpha_1 + \beta_1 y)/\sigma_1^2,$$

ρ^2 este mai direct interpretabil decît ρ însuși, fiind raportul dispersiei dreptei ajustate la dispersia totală. Pe de altă parte, dacă $\rho = 0$ nu implică independența, este dificil să interpretăm ρ ca măsură a independenței și este mai indicat să-l folosim ca un indicator decît ca o măsură precisă. Este recomandabil să se folosească ρ ca măsură a independenței numai în cazurile : normal sau aproape normal.

Scanned with OKEN Scanner

§ 4. Raport de corelație

Raportul de corelație remediază parțial defectele coeficientului de corelație, făcînd să intervină noțiunea de medie condiționată.

DEFINIȚIA 8. Raportul de corelație a lui X asupra lui Y este notat prin η_1 și este definit de

$$\eta_1^2 = D^2\{M(X \mid y)\}/\sigma_1^2. \tag{17}$$

Observație. η_1^2 nu este simetric în x și y.

DEFINIȚIA 8'. Raportul de corelație a lui Y asupra lui X este notat prin η_2^2 și este definit de :

$$\eta_2^2 = D^2\{M(Y \mid x)\}/\sigma_2^2 \tag{17'}$$

Observație. η_2 nu este simtric în y și x.

PROPOZIȚIA 5. *Dacă regresia este liniară, atunci* $\eta_1^2 = \rho^2$.

Demonstrație. Din (14), prin ridicare la pătrat și luarea mediei în raport cu y obținem

$$M\{M(X \mid y) - \mu_1)^2\} = D^2\{M(X \mid y)\} = \beta_1^2\sigma_2^2 = \mu_{11}^2/\sigma_2^2.$$

Din (9) și (17), rezultă propoziția

PROPOZIȚIA 6. $0 \leqslant \eta_1^2 \leqslant 1$.

Demonstrație. Avem :

$$\sigma_1^2 = M[\{X - M(X)\}^2] = M[(\{X - M(X \mid y)\} + \{M(X \mid y) - M(X)\})^2],$$

și cum

$$2M[\{X - M(X \mid y)\}\{M(X \mid y) - M(X)\}] =$$

$$= M_y[\{M(X \mid y) - M(X)\}M_{x \mid y}\{X - M(X \mid y)\}] = 0$$

putem scrie

$$\sigma_1^2 = M[\{X - M(X \mid y)\}^2] + M[\{M(X \mid y) - M(X)\}^2] =$$

$$= M[\{X - M(X \mid y)\}^2] + D^2\{M(X \mid y)\}. \tag{18}$$

Din (17) și (18) rezultă propoziția.

Observație. $\eta_1^2 = 1$ dacă și numai dacă

$$M[\{X - M(X \mid y)\}^2] = 0,$$

adică fiecare observație variază pe curba de regresie, astfel că X și Y sînt legate strict funcțional.

PROPOZIȚIA 7. *Au loc inegalitățile*

i) $\rho^2/\eta_1^2 \leqslant 1$,

ii) $0 \leqslant \rho^2 \leqslant \eta_1^2 \leqslant 1$ $(0 \leqslant \rho^2 \leqslant \eta_2^2 \leqslant 1)$.

Demonstrație. i) Putem scrie

$$\rho^2/\eta_1^2 = \mu_{11}^2/[\sigma_2^2 D^2\{M(X \mid y)\}]. \tag{19}$$

Scanned with OKEN Scanner

Din inegalitatea lui Schwarz

$$\mu_{11}^2 = (M_{X,Y}[\{Y - M(Y)\}\{X - M(X)\}])^2 =$$

$$= (M_Y[\{Y - M(Y)\}\{M(X|y) - M(X)\}])^2 \leqslant \tag{20}$$

$$\leqslant M_Y[\{Y - M(Y)\}^2] M_Y[\{M(X|y) - M(X)\}^2] = \sigma_2^2 D^2\{M(X|y)\},$$

egalitatea avînd loc dacă şi numai dacă

$$\{Y - M(Y)\} \text{ este proporţională cu } \{M(X|y) - M(X)\},$$

adică dacă $M(X|y)$ este o funcţie liniară strictă de y. Din (19) şi (20) rezultă i), egalitatea avînd loc numai cînd regresia lui X asupra lui Y este liniară.

ii) rezultă din propoziţia 6, punctul i).

În concluzie putem afirma

a) $\rho^2 = \eta_1^2 = 1$ dacă şi numai dacă X şi Y sînt într-o relaţie funcţională liniară strictă.

b) $\rho^2 \leqslant \eta_1^2 = 1$ dacă şi numai dacă X şi Y sînt într-o relaţie funcţională neliniară strictă.

c) $\rho^2 = \eta_1^2 < 1$ dacă şi numai dacă regresia lui X asupra lui Y este liniară

d) $\rho^2 < \eta_1^2 < 1$ implică că nu există nici o relaţie funcţională, şi o anumită regresie neliniară este o ajustare *mai bună* ca cea mai bună linie dreaptă $\{D^2(X|y)\} > D^2(\alpha_1 + \beta_1 Y)$.

PROPOZIŢIA 8. $\eta_1^2 = 1 \Leftrightarrow \eta_2^2 = 1$.

Observaţie. În general cele două rapoarte de corelaţie depăşesc ρ^2, însă după cum rezultă din exemplul următor $\eta_1^2 = \rho^2 < \eta_2^2$ dacă regresia lui X asupra lui Y este liniară în timp ce a lui Y asupra lui X nu este liniară.

Exemplul 6. În condiţiile exemplului 5 am văzut că regresia lui X asupra lui Y^2 este liniară cu coeficientul de regresie 0; astfel şi coeficientul de corelaţie dintre X şi Y^2 este 0. Deoarece am găsit că $M(X|y^2) = 0$, urmează de asemenea că $D\{M(X|y^2)\} = 0$ astfel că raportul de corelaţie a lui X asupra lui Y^2 este 0 (deoarece coeficientul de corelaţie este zero şi regresia liniară).

Regresia lui Y^2 asupra lui X nu este liniară: în exemplul 3 am găsit că

$$M(Y^2|x) = 1 + \rho^2(x^2 - 1)$$

astfel încît $D^2\{M(Y^2|x)\} = M[\{\rho^2(x^2 - 1)\}^2] = \rho^4 M[\{x^2 - 1\}^2] = 2\rho^4$ şi $\sigma_2^2 = 2$, deci raportul de corelaţie a lui Y^2 asupra lui X este ρ^4, care întotdeauna depăşeşte coeficientul de corelaţie dintre X şi Y^2, care este zero, cînd $\rho \neq 0$.

§ 5. Legătura dintre curba de regresie şi cumulanţi

Putem scrie

$$f(x, y) = g(x)h_x(y), \tag{21}$$

unde $g(x)$ este densitatea de repartiţie marginală a variabilei X şi $h_x(y)$ este densitatea de repartiţie a variabilei Y condiţionate de $X = x$,

Scanned with OKEN Scanner

sau în cazul în care avem p regresori $X_1, X_2, \ldots, X_p$,

$$f(y, x_1, \ldots, x_p) = g(x)h_x(y) \tag{21'}$$

unde $g(x)$ este densitatea de repartiţie marginală a variabilelor $X_1, X_2, \ldots, X_p$.

Funcţia caracteristică comună variabilelor X şi Y este

$$\varphi(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(it_1x + it_2y)g(x)h_x(y)\,dx\,dy =$$

$$= \int_{-\infty}^{\infty} \exp(it_1x)g(x)\varphi_x(t_2)\,dx, \tag{22}$$

unde

$$\varphi_x(t_2) = \int_{-\infty}^{\infty} \exp(it_2y)h_x(y)\,dy$$

este funcţia caracteristică a variabilei Y condiţionate de $X = x$. Avem

$$\mu'_{rx} = \frac{1}{i^r}\left[\frac{\partial^r}{\partial t_2^r}\varphi_x(t_2)\right]_{t_2=0} \tag{23}$$

şi deci din (22) şi (23)

$$\left[\frac{\partial^r}{\partial t_2^r}\varphi(t_1, t_2)\right]_{t_2=0} = i^r\int_{-\infty}^{\infty} \exp(it_1x)g(x)\mu'_{rx}\,dx. \tag{24}$$

Din teorema de inversiune, găsim

$$g(x)\mu'_{rx} = \frac{(-i)^r}{2\pi}\int_{-\infty}^{\infty} \exp(-it_1x)\left[\frac{\partial^r}{\partial t_2^r}\varphi(t_1, t_2)\right]_{t_2=0} dt_1. \tag{25}$$

Din (25) se poate scrie regresia de orice ordin. Astfel pentru $r = 1$, obţinem

$$g(x)\mu'_{1x} = \frac{-1}{2\pi}\int_{-\infty}^{\infty} \exp(-it_1x)\left[\frac{\partial}{\partial t_2}\varphi(t_1, t_2)\right]_{t_2=0} dt_1. \tag{26}$$

Funcţia caracteristică comună a variabilelor $Y, X_1, \ldots, X_p$ este

$$\varphi(u, t_1, \ldots, t_p) = \int\ldots\int \exp\left(iuy + i\sum_{j=1}^{p} t_jx_j\right)g(x)h_x(y)\,dx\,dy =$$

$$= \int\ldots\int \exp\left(i\sum_{j=1}^{p} t_jx_j\right)g(x)\varphi_x(u)\,du.$$

Scanned with OKEN Scanner

Avem

$$\mu'_{rx} = \left[\frac{\partial^r}{\partial u^r}\,\varphi_x(u)\right]_{u=0} \tag{23'}$$

$$\left[\frac{\partial^r}{\partial u^r}\,\varphi(u, t_1, \ldots, t_p)\right]_{u=0} = \mathrm{i}^r \int \exp(\mathrm{i}\sum t_j x_j) g(x) \mu'_{rx}\,\mathrm{d}x \tag{24'}$$

$$g(x)\mu'_{rx} = \frac{(-\mathrm{i})^r}{2\pi}\int \exp(-\mathrm{i}\sum t_j x_j)\left[\frac{\partial^r}{\partial u^r}\,\varphi(u, t_1, \ldots, t_p)\right]_{u=0}\mathrm{d}t \tag{25'}$$

Dacă există toţi cumulanţii, atunci

$$\varphi(t_1, t_2) = \exp\left\{\sum_{r,s=0}^{\infty} \mathfrak{K}_{rs}\frac{(\mathrm{i}t_1)^r}{r!}\cdot\frac{(\mathrm{i}t_2)^s}{s!}\right\},$$

unde $\mathfrak{K}_{00}$ este definit ca fiind egal cu zero. Deci

$$\left[\frac{\partial\varphi(t_1, t_2)}{\partial t_2}\right]_{t_2=0} = \left[\mathrm{i}\varphi(t_1, t_2)\sum_{r=0}^{\infty}\sum_{s=1}^{\infty}\mathfrak{K}_{rs}\frac{(\mathrm{i}t_1)^r}{r!}\cdot\frac{(\mathrm{i}t_2)^{s-1}}{(s-1)!}\right]_{t_2=0} = \tag{27}$$

$$= \mathrm{i}\varphi(t_1, 0)\sum_{r=0}^{\infty}\mathfrak{K}_{r1}\frac{(\mathrm{i}t_1)^r}{r!}.$$

Ţinînd seama în (26) de (27), obţinem

$$g(x)\mu'_{1x} = \frac{1}{2\pi}\int_{-\infty}^{\infty}\exp(-\mathrm{i}t_1 x)\varphi(t_1, 0)\sum_{r=0}^{\infty}\mathfrak{K}_{r1}\frac{(\mathrm{i}t_1)^r}{r!}\mathrm{d}t_1 \tag{28}$$

şi dacă este permisă integrarea termen cu termen a seriei, (28) devine

$$g(x)\mu'_{1x} = \frac{1}{2\pi}\sum_{r=0}^{\infty}\frac{\mathfrak{K}_{r1}}{r!}\mathrm{i}^r\int_{-\infty}^{\infty}t_1^r\exp(-\mathrm{i}t_1 x)\varphi(t_1, 0)\,\mathrm{d}t_1. \tag{29}$$

După teorema de inversiune

$$g(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\exp(-\mathrm{i}t_1 x)\varphi(t_1, 0)\,\mathrm{d}t_1$$

şi din condiţiile de existenţă :

$$(-D)^j g(x) \equiv (-1)^j\frac{\mathrm{d}^j}{\mathrm{d}x^j}g(x) = \frac{\mathrm{i}^j}{2\pi}\int_{-\infty}^{\infty}t_1^j\exp(-\mathrm{i}t_1 x)\varphi(t_1, 0)\,\mathrm{d}t_1. \tag{30}$$

Scanned with OKEN Scanner

Folosind (30), (29) devine

$$g(x)\mu'_{1x} = \sum_{r=0}^{\infty} \frac{\mathcal{K}_{r1}}{r!}(-D)^r g(x). \tag{31}$$

Astfel pentru regresia lui Y asupra lui X, avem

$$\mu'_{1x} = \sum_{r=0}^{\infty} \frac{\mathcal{K}_{r1}}{r!} \frac{(-D)^r g(x)}{g(x)}. \tag{32}$$

Deci am demonstrat teorema:

TEOREMA 1. (Wicksell (1934)) *Dacă toți cumulanții există și dacă este permisă integrarea termen cu termen a seriei din* (28) (*în particular dacă* $g(x)$ *și toate derivatele sale sînt continui în domeniul de variație a lui* x *și* 0 *în extremitățile sale*), *atunci are loc* (32).

Exemplul 7. Dacă $g(x)$ este $n(x; 0, 1)$, (32) devine

$$\mu'_{1x} = \sum_{r=0}^{\infty} \frac{\mathcal{K}_{r1}}{r!} H_r(x), \tag{33}$$

unde H_r este polinomul Cebîșev—Hermite de ordinul r.

Exemplul 8. Pentru repartiția normală bidimensională, funcția caracteristică a variabilelor $\frac{X-\mu_1}{\sigma_1}$ și $\frac{Y-\mu_2}{\sigma_2}$ este

$$\varphi(t_1, t_2) = \exp\left\{-\frac{1}{2}(t_1^2 + t_2^2 + 2\rho t_1 t_2)\right\},$$

de unde

$$\mathcal{K}_{01} = 0, \quad \mathcal{K}_{r1} = 0, \quad r > 1,$$

astfel încît

$$\mathcal{K}_{11} = \rho$$

este singurul cumulant nenul din (32). Repartiția marginală $g(x)$ este normală redusă, astfel încît (32) devine (33) și avem

$$\mu'_{1x} = \mathcal{K}_{11} H_1(x) = \rho x.$$

Aceasta este regresia variabilei $(Y-\mu_2)/\sigma_2$ asupra variabilei $(X-\mu_1)/\sigma_1$. Urmează că regresia variabilei Y asupra variabilei X este

$$M(Y \mid x) - \mu_2 = \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1).$$

Exemplul 9. Dacă (X_i, Y_i), $1 \leqslant i \leqslant n$ este o selecție asupra unui vector normal bidimensional, atunci funcția caracteristică comună a

variabilelor

$$U = \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu_1)^2/\sigma_1^2 \quad \text{și} \quad V = \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu_2)^2/\sigma_2$$

este

$$\varphi(t_1, t_2) = \{(1 - \theta_1)(1 - \theta_2) - \rho^2\theta_1\theta_2\}^{-(1/2)n} \tag{34}$$

unde $\theta_1 = it_1$, $\theta_2 = it_2$. Avem

$$\left[\frac{\partial^r \varphi(t_1, t_2)}{\partial t_2^r}\right]_{t_2=0} = i^r\left[\frac{\partial^r \varphi(t_1, t_2)}{\partial \theta_2^r}\right]_{\theta_2=0} =$$

(35)

$$= i^r\left(\frac{1}{2}n + r - 1\right)^{(r)} \frac{\{1 - (1 - \rho^2)\theta_1\}^r}{(1 - \theta_1)^{(n/2)+r}}.$$

Astfel din (25) și (35),

$$g(u)\mu'_{rv} = \left(\frac{1}{2}n + r - 1\right)^{(r)} \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-\theta_1 u) \frac{\{\rho^2 + (1 - \rho^2)(1 - \theta_1)\}^r}{(1 - \theta_1)^{(n/2)+r}}\, dt_1. \tag{36}$$

Avem :

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\exp(-\theta_1 u)}{(1 - \theta_1)^k}\, dt_1 = \frac{1}{\Gamma(k)} e^{-u} u^{k-1},$$

$$g(u) = \frac{1}{\Gamma\left(\frac{1}{2}n\right)} e^{-u} u^{(n/2)-1}.$$

Ținînd seama de acestea în (36), punînd $r = 1, 2$, găsim

$$\mu'_{1v} = \frac{1}{2}n\left\{\rho^2 \frac{v}{\left(\frac{1}{2}n\right)} + (1 - \rho^2)\right\} = \rho^2 v + \frac{1}{2}n(1 - \rho^2) \tag{37}$$

și respectiv

$$\mu'_{2v} = \frac{1}{2}n\left(\frac{1}{2}n + 1\right)\left\{\rho^4 \frac{v^2}{\left(\frac{1}{2}n\right)\left(\frac{1}{2}n + 1\right)} + 2\rho^2(1 - \rho^2)\frac{v}{\frac{1}{2}n} + (1 - \rho^2)^2\right\} =$$

$$= \rho^4 x^4 + 2\rho^2(1 - \rho^2)v\left(\frac{1}{2}n + 1\right) + (1 - \rho^2)^2 \frac{1}{n}n\left(\frac{1}{2}n + 1\right),$$

astfel încît

$$\mu_{2v} = \mu'_{2v} - (\mu'_{1v})^2 = (1 - \rho^2)\left(2\rho^2 v + \frac{1}{2} n(1 - \rho^2)^2\right). \tag{38}$$

Observație. (37) și (38) arată că regresiile mediei și dispersiei lui U asupra lui V sînt liniare.

TEOREMA 2. *Dacă regresia lui Y asupra lui X este liniară atunci*

$$\mu'_{1x} = M(Y \mid x) = \beta_0 + \beta_1 x, \tag{39}$$

atunci

$$\left[\frac{\partial \ln \varphi(t_1, t_2)}{\partial t_2}\right]_{t_2=0} = i\beta_0 + \beta_1 \frac{\partial \ln \varphi(t_1, 0)}{\partial t_1} \tag{40}$$

și reciproc, dacă repartiția marginală $g(x)$ este completă, (40) *este suficientă și necesară pentru* (39).

Demonstrație. Din (24), cu $r = 1$, folosind (39) avem

$$\left[\frac{\partial \varphi(t_1, t_2)}{\partial t_2}\right]_{t_2=0} = i\int_{-\infty}^{\infty} \exp(it_1 x) g(x)(\beta_0 + \beta_1 x)\, dx = \tag{41}$$

$$= i\beta_0 \varphi(t_1, 0) + \beta_1 \frac{\partial}{\partial t_1} \varphi(t_1, 0). \tag{42}$$

Dacă în (42) punem $\psi(t_1, t_2) = \log \varphi(t_1, t_2)$ și împărțim prin $\varphi(t_1, 0,)$ obținem (40). Reciproc, dacă are loc (40), îl putem rescrie cu ajutorul lui (24) în forma

$$i\int_{-\infty}^{\infty} \exp(it_1 x)(\beta_0 + \beta_1 x - \mu'_{1x}) g(x)\, dx = 0. \tag{43}$$

(43) implică că $\exp(it_1 x)(\beta_0 + \beta_1 x - \mu'_{1x}) = 0$ identic în x dacă $g(x)$ este completă și deci (39) urmează.

Observații. i). Dacă există toți cumulanții și dacă folosim (27), (40) dă

$$\sum_{r=0}^{\infty} \varkappa_{r1} \frac{(it_1)^r}{r!} = \beta_0 + \beta_1 \sum_{r=0}^{\infty} \varkappa_{r0} \frac{(it_1)^{r-1}}{(r-1)!}. \tag{44}$$

Identificînd coeficienții lui t din (44), obținem

$$(r = 0)\quad \varkappa_{01} = \beta_0 + \beta_1 \varkappa_{10} \tag{45}$$

$$(r \geqslant 1)\quad \varkappa_{r1} = \beta_1 \varkappa_{r+1,0}. \tag{46}$$

Condiția (46) pentru liniaritatea regresiei este dată de Wicksell (1934). (45) și (46) sînt suficiente, cît și necesare, pentru (40) și de aici (dată completitudinea lui $g(x)$) pentru (39).

Scanned with OKEN Scanner

ii) Dacă în loc să exprimăm (40) cu ln φ, îl exprimăm cu φ, ca în (42), găsim analogul lui (46) pentru momente cetrate

$$\mu_{r1} = \beta_1 \mu_{r+1,0}. \tag{47}$$

Dacă regresia lui X asupra lui Y este de asemenea liniară, de forma

$$x = \beta_0' + \beta_1' y,$$

avem

$$\mathcal{K}_{1r} = \beta_1' \mathcal{K}_{0,r+1}, \quad r \geqslant 1. \tag{48}$$

Cînd $r = 1$, (46) și (48) dau

$$\mathcal{K}_{11} = \beta_1 \mathcal{K}_{20} = \beta_1' \mathcal{K}_{02},$$

de unde

$$\beta_1 \beta_1' = \mathcal{K}_{11}^2 / (\mathcal{K}_{20} \mathcal{K}_{02}) = \rho^2, \tag{49}$$

ρ fiind coeficientul de corelație dintre X și Y.

TEOREMA 3. *Dacă regresia lui Y asupra lui $X_1, \ldots, X_p$ este liniară, adică*

$$M(Y \mid x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \tag{50}$$

atunci

$$\left[\frac{\partial \psi(u, t_1, \ldots, t_p)}{\partial u}\right]_{u=0} = i\beta_0 + \sum_{j=1}^{p} \beta_j \frac{\partial}{\partial t_j} \psi(0, t_1, \ldots, t_p) \tag{51}$$

și reciproc

Observație. În cazul a p regresori $X_1, X_2, \ldots, X_p$, (46) devine

$$\mathcal{K}_{1,r_1,r_2,\ldots,r_p} = \beta_1 \mathcal{K}_{0,r_1+1,r_2,\ldots,r_p} + \beta_2 \mathcal{K}_{0,r_1,r_2+1,r_3,\ldots,r_p} + \beta_p \mathcal{K}_{0,r_1,\ldots,r_{p-1},r_p+1}. \tag{52}$$

Presupunem că numai media lui Y se schimbă cu x. Ne vom referi la această restricție spunînd că Y are erori identice. Astfel există ε astfel încît

$$y = \mu_{1x}' + \varepsilon. \tag{53}$$

În particular, dacă regresia este liniară, (53) este

$$y = \beta_0 + \beta_1 x + \varepsilon. \tag{54}$$

Dacă y are erori identice, (21) devine

$$f(x, y) = g(x) h(\varepsilon). \tag{55}$$

Reciproc (55) implica erori identice pentru y. Rezultatul corespunzător pentru funcția caracteristică este

TEOREMA 4. *Dacă regresia lui Y asupra lui X este liniară cu erori identice, atunci*

$$\varphi(t_1, t_2) = \varphi_g(t_1 + t_2 \beta_1) \varphi_h(t_2) \exp(i t_2 \beta_0). \tag{56}$$

Demonstrație. Avem :

$$\varphi(t_1, t_2) = \iint \exp(it_1x + it_2y)f(x, y)\,dx\,dy =$$

$$= \iint \exp[it_1x + it_2(\beta_0 + \beta_1x + \varepsilon)]g(x)h(\varepsilon)\,dx\,dy =$$

$$= \int \exp[i(t_1 + t_2\beta_1)x]g(x)\,dx \int \exp(it_2\varepsilon)h(\varepsilon)\,d\varepsilon(\exp(it_2\beta_0)),$$

ceea ce demonstrează teorema.

TEOREMA 5 (Bernstein (1928), Féron, Fourgeaud (1952), Kendall, Stuart (1961)). *Dacă regresiile lui Y asupra lui X și a lui X asupra lui Y sînt liniare cu erori identice, atunci exceptînd cazurile :*

a) *X și Y sînt independente sau*

b) *X și Y sînt legate funcțional,*

X și Y sînt repartizate normal bidimensional.

Demonstrație. Din (56) prin logaritmare, obținem

$$\psi(t_1, t_2) = \psi_g(t_1 + t_2\beta_1) + \psi_h(t_2) + it_2\beta_0. \tag{57}$$

și similar din regresia lui X asupra lui Y

$$\psi(t_1, t_2) = \psi_{g'}(t_2 + t_1\beta_1') + \psi_{h'}(t_1) + it_1\beta_0' \tag{58}$$

unde semnul „'" este folosit pentru a distinge coeficienții și repartițiile de acelea din (57). Egalînd (57) și (58) și considerînd succesiv puterile lui t_1 și t_2, notînd cumulantul de ordin r al lui g prin $\varkappa_{r0}$, acela pentru g' prin $\varkappa_{0r}$, acela pentru h prin λ_{r0} și acela pentru h' prin λ_{02} :

Prima putere

$$\varkappa_{01}i(t_1 + t_2\beta_1) + \lambda_{10}it_2 + it_2\beta_0 = \varkappa_{01}i(t_2 + t_1\beta_1') + \lambda_{01}it_1 + it_1\beta_0'.$$

sau egalînd coeficienții lui t_1 și ai lui t_2,

$$\varkappa_{10} = \varkappa_{01}\beta_1' + \lambda_{01} + \beta_0' \tag{59}$$

$$\varkappa_{10}\beta_1 + \lambda_{10} + \beta_0 = \varkappa_{01}. \tag{60}$$

Putem presupune că erorile au medii zero. Dacă măsurăm x și y de la mediile lor, (59) și (60) dau

$$\beta_0' = \beta_0 = 0. \tag{61}$$

A doua putere

$$\varkappa_{20}(t_1 + t_2\beta_1)^2 + \lambda_{20}t_2^2 = \varkappa_{02}(t_2 + t_1\beta_1')^2 + \lambda_{02}t_1^2,$$

din care, prin egalarea coeficienţilor lui t_1^2, t_1t_2 şi t_2^2, obţinem :

$$\mathcal{K}_{20} = \mathcal{K}_{02}(\beta_1')^2 + \lambda_{02}, \tag{62}$$

$$\mathcal{K}_{20}\beta_1 = \mathcal{K}_{02}\beta_1', \tag{63}$$

$$\mathcal{K}_{20}\beta_1^2 + \lambda_{20} = \mathcal{K}_{02} \tag{64}$$

(62)—(64) dau relaţiile între g, h, g' şi h'; în particular (63) dă raportul β_1/β_1' ca fiind egal cu $\mathcal{K}_{02}/\mathcal{K}_{20}$.

A treia putere

$$\mathcal{K}_{30}\{i(t_1 + t_2\beta_1)\}^3 + \lambda_{30}(it_2)^3 = \mathcal{K}_{03}\{i(t_2 + t_1\beta_1')\}^3 + \lambda_{03}(it_1)^3$$

Termeni în $t_1^2t_2$ şi $t_1t_2^2$ dau

$$\mathcal{K}_{30}\beta_1 = \mathcal{K}_{03}(\beta_1')^2 \tag{65}$$

$$\mathcal{K}_{30}\beta_1^2 = \mathcal{K}_{03}\beta_1'. \tag{66}$$

Lăsînd la o parte posibilităţile β_1, $\beta_1' = 0$ sau $\beta_1\beta_1' = 1$, (65) şi (66) implică $\mathcal{K}_{30} = \mathcal{K}_{03} = 0$. Similar dacă luăm puterea a patra şi puterea a cincea găsim că toţi cumulanţii $\mathcal{K}_{r0}$, $\mathcal{K}_{0r}$ de ordin superior trebuie să se anuleze. Din ecuaţiile

$$\mathcal{K}_{30} = \mathcal{K}_{03}(\beta_1')^3 + \lambda_{03},$$

$$\mathcal{K}_{30}\beta_1^3 + \lambda_{30} = \mathcal{K}_{03},$$

urmează că cumulanţii trebuie să se anuleze. Astfel toate repartiţiile g, h, g', h' sînt normale şi din (57) şi (58) urmează că $\psi(t_1, t_2)$ este pătratică în t_1, t_2 şi deci că (X, Y) este un vector normal bidimensional. În cazurile excepţionale, pe care le-am neglijat,această afirmaţie nu mai este adevărată. Dacă $\beta_1\beta_1' = 1$, corelaţia dintre X şi Y, după (49), este ± 1 şi X este o funcţie strict liniară de Y; dacă pe de altă parte β_1 sau $\beta_1' = 0$, variabilele X şi Y sînt independente. Teorema este demonstrată.

TEOREMA 6. *Dacă fiecare din regresiile liniare ale mulţimii de p variabilele are erori identice, variabilele sînt repartizate normal multidimensional.*

§ 6. Regresii liniare parţiale

Considerăm extensia relaţiilor regresiei liniare din § 2 la p variabile. Avem (în cazul normal) :

$$\frac{M(X_1 \mid x_2, \ldots, x_p)}{\sigma_1} = -\sum_{j=2}^{p} \frac{C_{1j}}{C_{11}} \cdot \frac{x_j}{\sigma_j}. \tag{67}$$

Vom nota coeficientul de regresie a lui X_1 asupra lui X_j cu celelalte $p - 2$ variabile fixate prin

$$\beta_{1j.2.3\ldots j-1,\,j+1,\ldots,p} \text{ sau mai scurt } \beta_{1j.q_j}$$

unde q semnifică pentru „celelalte variabile care nu sînt indici primari" : indicile la q este pentru a distinge între diferiţii q.

DEFINIŢIA 9. $\beta_{1j.q}$ poartă numele de coeficienţii de regresie parţiali.

Scanned with OKEN Scanner

Prin urmare, avem

$$M(X_1 \mid x_2, \ldots, x_p) = \beta_{12 \cdot q_2} x_2 + \beta_{13 \cdot q_3} x_3 + \ldots + \beta_{1p \cdot q_p} x_p. \tag{68}$$

În cazul normal multidimensional compararea lui (68) și (67) dă

$$\beta_{1j \cdot q_j} = -\frac{\sigma_1}{\sigma_j} \cdot \frac{C_{1j}}{C_{11}}. \tag{69}$$

Similar, coeficientul de regresie al lui x_j asupra lui x_1 cu celelalte variabile fixate este

$$\beta_{j1 \cdot q_j} = -\frac{\sigma_j}{\sigma_1} \cdot \frac{C_{j1}}{C_{jj}} \tag{70}$$

și astfel, deoarece $C_{1j} = C_{j1}$, (XXVIII.186) (69) și (70) dau

$$\rho^2_{1j \cdot q_j} = -\frac{C^2_{1j}}{C_{11} C_{jj}} = \beta_{1j \cdot q_j} \beta_{j1 \cdot q_j} \tag{71}$$

(69), (70) constituie expresii generale ale coeficienților de regresie.

Observație. $\beta_{1j \cdot q_j}$ nu este simetric în x_1 și x_j

Definiția 10. Statistica

$$x_{12 \ldots p} = X_1 - M(X_1 \mid x_2, \ldots, x_p) \tag{72}$$

poartă numele de rezidul de ordin $(p - 1)$.

Propoziția 9. *Au loc relațiile*

$$M(x_{1.2 \ldots p}) = 0$$

și

$$\sigma^2_{1.2 \ldots p} = M(x^2_{1.2 \ldots p}) = M[\{X_1 - M(X_1 \mid x_2, \ldots, x_p)\}^2].$$

Observație. $\sigma^2_{1.2 \ldots p}$ este dispersia rezidului lui X_1 față de regresie.

Propoziția 10. *Avem*

$$\sigma^2_{1.2 \ldots p} = M\left\{X_1^2 - \sum_{j=2}^{p} \beta_{1j \cdot q_j} X_1 X_j\right\} = \sigma_1^2 - \sum_{j=2}^{p} \beta_{1j \cdot q_j} \sigma_{1j}. \tag{73}$$

Demonstrație. Din (68), obținem

$$\sigma^2_{1.2 \ldots p} = M\left[\left\{X_1 - \sum_{j=2}^{p} \beta_{1j \cdot q_j} X_j\right\}^2\right] \tag{74}$$

sau

$$\sigma^2_{1.2 \ldots p} = M\left[X_1\left(X_1 - \sum_{j=2}^{p} \beta_{1j \cdot q_j} X_j\right) - \sum_{j=2}^{p} \beta_{1j \cdot q_j} X_j\left(X_1 - \sum_{j=2}^{p} \beta_{1j \cdot q_j} X_j\right)\right] \tag{75}$$

Fixînd $x_2, \ldots, x_p$, găsim că media condiționată a celui de al doilea produs din membrul drept al relației (75) este, după (68), zero, ceea ce demonstrează propoziția.

Observație. i) Dispersia rezidului (73) este independentă de valorile fixate pentru $x_2, \ldots, x_p$ dacă $\beta_{1j \cdot q_j}$ sînt independenți de aceste valori.

Scanned with OKEN Scanner

ii) În cazul normal, dispersiile și covarianțele condiționate obținute fixînd o mulțime de valori ale variabilelor nu depind de valorile fixate.

PROPOZIȚIA 11. *Avem*:

$$\sigma^2_{1.2\ldots p}/\sigma^2_1 = \frac{|\mathbf{C}|}{C_{11}}.$$

Demonstrație. Din (73) și (69) găsim

$$\sigma^2_{1.2\ldots p} = \sigma^2_1 + \sum_{j=2}^{p} \frac{\sigma_1}{\sigma_j} \cdot \frac{C_{1j}}{C_{11}} \sigma_{1j} \tag{76}$$

sau

$$\sigma^2_{1.2\ldots p}/\sigma^2_1 = 1 + \frac{1}{C_{11}} \sum_{j=2}^{p} C_{1j}\rho_{1j} = 1 + \frac{1}{C_{11}}(|\mathbf{C}| - C_{11}) = \frac{|\mathbf{C}|}{C_{11}}, \tag{77}$$

ceea ce demonstrează propoziția.

Observație. (77) se mai poate scrie.

$$\sigma^2_{1\cdot q} = \sigma^2_1 \frac{|\mathbf{C}|}{C_{11}} \tag{78}$$

și similar dacă 1 este înlocuit prin orice alt indice.

PROPOZIȚIA 12. *Avem*:

$$\operatorname{cov}(x_{l\cdot q_l}, x_{m\cdot q_m}) = \sigma_l \sigma_m \frac{|\mathbf{C}|}{C_{lm}}, \tag{79}$$

DEFINIȚIA 11. *Coeficienții de regresie sînt*

$$\operatorname{cov}(x_{lr}, x_{mr})/\sigma^2_{mr} = \beta_{lm\cdot r}, \tag{80}$$

$$\operatorname{cov}(x_{lr}, x_{mr})/\sigma^2_{lr} = \beta_{ml\cdot r}, \tag{81}$$

$$\operatorname{cov}(x_{lr}, x_{mr})/(\sigma_{lr}\sigma_{mr}) = \rho_{lmr} \tag{82}$$

COROLARUL 1. *Avem*

$$\sigma^2_{1.q-2} = \sigma^2_1 \frac{|\mathbf{D}|}{D_{11}} \tag{83}$$

$$\sigma^2_{1.q} = \sigma^2_1 \cdot \frac{|\mathbf{C}|}{C_{11}}, \tag{84}$$

unde $|\mathbf{D}|$ *este determinantul corelațiilor pentru toate variabilele exceptînd* x_2 *iar indicile* $q - 2$ *reprezintă mulțimea* q *exceptînd* x_2.

Scanned with OKEN Scanner

Demonstrația rezultă din (78).

Observație. Avem

$$\frac{\sigma^2_{1q}}{\sigma^2_{1.q-2}} = \frac{D_{11}}{C_{11}} \frac{|\mathbf{C}|}{|\mathbf{D}|}. \tag{85}$$

Prin definiție $|\mathbf{D}| = C_{22}$ și

$$\begin{vmatrix} C_{11} & C_{12} \\ C_{12} & C_{22} \end{vmatrix} = |\mathbf{C}| D_{11}. \tag{86}$$

Folosind (86), (85) devine

$$\frac{\sigma^2_{1.q}}{\sigma^2_{1.q-2}} = \frac{\begin{vmatrix} C_{11} & C_{12} \\ C_{12} & C_{22} \end{vmatrix}}{C_{11} \; C_{22}} = 1 - \frac{C^2_{12}}{C_{11} C_{22}} \tag{87}$$

sau, ținînd seama de (XXVIII.186),

$$\sigma^2_{1.q} = \sigma^2_{1.q-2}(1 - \rho^2_{12.q}). \tag{88}$$

Din aproape în aproape

$$\sigma^q_{1.q-2} = \sigma^2_{1.q-2-3}(1 - \rho^2_{13.q-2}),$$

$$\sigma^2_{1.2\ldots p} = \sigma^2_1(1 - \rho^2_{1p})(1 - \rho^2_{1(p-1).p})(1 - \rho^2_{1(p-2)\,(p-1)p}) \ldots (1 - \rho^2_{12\ldots p}) \tag{89}$$

(89) se mai poate scrie

$$\frac{\sigma^2_{1.2\ldots p}}{\sigma^2_1} = (1 - \rho^2_{12})(1 - \rho^2_{13.2})(1 - \rho^2_{14.23}) \ldots (1 - \rho^2_{1p.23\ldots(p-1)}) = \frac{|\mathbf{C}|}{C_{11}}. \tag{90}$$

Propoziția 13. *Au loc relațiile*

$$\beta_{12\cdot pr} = \frac{\beta_{12\cdot r} - \beta_{1p\cdot r}\beta_{p2\cdot r}}{1 - \beta_{2p\cdot r}\beta_{p2\cdot r}} \tag{91}$$

$$\rho_{12\cdot pr} = (\beta_{12\cdot pr}\beta_{21\cdot pr})^{1/2} = \frac{\rho_{12\cdot r} - \rho_{1p\cdot r}\rho_{2pr}}{\{(1 - \rho^2_{1p.r})(1 - \rho^2_{2p.r})\}^{1/2}}. \tag{92}$$

Demonstrație. Din (XXVIII.195), covarianța dintre x_1, x_2 cu x_p fixat, este

$$\sigma_{12\cdot p} = \sigma_{12} - \sigma_{1p}\sigma_{p2}/\sigma^2_p$$

și dacă adăugăm indicii $3, \ldots, (p-1)$, avem

$$\sigma_{12\cdot 3\cdots p} = \sigma_{12\cdot 3\cdots(p-1)} - \frac{\sigma_{1p\cdot 3\cdots(p-1)}\sigma_{p2\cdot 3\cdots(p\cdot 1)}}{\sigma^2_{p.3\ldots(p-1)}}. \tag{93}$$

Scanned with OKEN Scanner

Folosind definiţia 11 a coeficienţilor de regresie ca rapoarte de covarianţe la dispersii, adică

$$\beta_{ij\cdot k} = \frac{\sigma_{ij\cdot k}}{\sigma^2_{j.k}}$$

din (93) cu $3, \ldots, (p-1)$ înlocuiţi prin r, găsim

$$\beta_{12\cdot pr}\,\sigma^2_{2.pr} = \beta_{12\cdot r}\,\sigma^2_{2.r} - \beta_{1p\cdot r}\beta_{p2\cdot r}\,\sigma^2_{2r}$$

sau

$$\beta_{12\cdot pr} = \frac{\sigma^2_{2.r}}{\sigma^2_{2pr}}(\beta_{12\,r} - \beta_{1p\cdot r}\beta_{p2\cdot r}). \tag{94}$$

Dacă în (94) punem $x_1 \equiv x_2$, obţinem

$$\sigma^2_{2\cdot pr} = \sigma^2_{2\cdot r}(1 - \beta_{2p\cdot r}\beta_{p2\,r}) = \sigma^2_{2\cdot r}(1 - \rho^2_{2p\cdot r}), \tag{95}$$

o altă formă a lui (88). Din (94) şi (95), obţinem (91).

Din (71) şi (91) găsim (92).

§ 7. Coeficienţii de regresie de selecţie

DEFINIŢIA 12. Statisticile

$$\begin{cases} b_1 = m_{11}/s_2^2 = \dfrac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y}) \Big/ \dfrac{1}{n}\sum(y_i - \bar{y})^2 \\ b_2 = m_{11}/s_1^2 = \dfrac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y}) \Big/ \dfrac{1}{n}\sum(x_i - \bar{x})^2 \end{cases} \tag{96}$$

poartă numele de coeficienţi de regresie de selecţie.

Pentru uşurinţa calculelor se folosesc expresiile

$$\begin{cases} b_1 = \dfrac{\sum x_i y_i - (\sum x_i)(\sum y_i)/n}{\sum y_i^2 - (\sum y_i)^2/n} \\ b_2 = \dfrac{\sum x_i y_i - (\sum x_i)(\sum y_i)/n}{\sum x_i^2 - (\sum x_i)^2/n} \end{cases} \tag{97}$$

Cum

$$D^2\left(\frac{x_1}{x_2}\right) = \frac{D^2(x_1)}{\mu_1^2} + \frac{\mu_1^2 D^2(x_2)}{\mu_2^4} - \frac{2\mu_1 \operatorname{cov}(x_1, x_2)}{\mu_2^3} =$$

$$= \left\{\frac{M(x_1)}{M(x_2)}\right\}^2 \left\{\frac{D^2(x_1)}{M^2(x_1)} + \frac{D^2(x_2)}{M^2(x_2)} - \frac{2\operatorname{cov}(x_1, x_2)}{M(x_1)M(x_2)}\right\},$$

Scanned with OKEN Scanner

obținem

$$D^2(b_2) = \left(\frac{\mu_{11}}{\sigma_1^2}\right)^2 \left\{\frac{D^2(m_{11})}{\mu_{11}^2} + \frac{D^2(s_1^2)}{\sigma_1^4} - \frac{2\,\mathrm{cov}(m_{11}, s_1^2)}{\mu_{11}\sigma_1^2}\right\}.$$

Ținînd seama de teorema II.13 care dă expresii pentru dispersii și covarianțe, găsim

$$D^2(b_2) = \frac{1}{n}\left(\frac{\mu_{11}}{\sigma_1^2}\right)^2 \left\{\frac{\mu_{22}}{\mu_{11}^2} + \frac{\mu_{40}}{\sigma_1^4} - \frac{2\mu_{31}}{\mu_{11}\sigma_1^2}\right\}. \tag{99}$$

Prin urmare am demonstrat

PROPOZIȚIA 14. *Dispersia statisticii b_2 este dată de* (99).

PROPOZIȚIA 15. *Dacă (X_1, X_2) este un vector normal bidimensional avem*

$$D^2(b_2) = \frac{1}{n}\frac{\sigma_2^2}{\sigma_1^2}(1 - \rho^2). \tag{100}$$

Demonstrație. Dacă în (99) ținem seama de propoziția XXVIII.4' găsim :

$$D^2(b_2) = \frac{1}{n}\left(\frac{\mu_{11}}{\sigma_1^2}\right)^2 \left\{\frac{(1 + 2\rho^2)\,\sigma_1^2\sigma_2^2}{\mu_{11}^2} + \frac{3\sigma_1^2}{\sigma_1^4} - \frac{6\rho\sigma_1^3\sigma_2}{\mu_{11}\sigma_1^2}\right\} =$$

$$= \frac{1}{n}\left(\frac{\mu_{11}}{\sigma_1^2}\right)^2 \left\{\frac{\sigma_1^2\sigma_2^2}{\mu_{11}^2} - 1\right\},$$

ceea ce demonstrează propoziția.
Similar

$$D^2(b_1) = \frac{1}{n}\frac{\sigma_1^2}{\sigma_2^2}(1 - \rho^2). \tag{101}$$

TEOREMA 7. *În selecții din populații normale, b_2 are probabilitatea elementară*

$$\mathrm{d}F = \frac{\Gamma\left(\frac{n}{2}\right)\sigma_2^{n-1}(1-\rho^2)^{(n-1)/2}}{\sqrt{\pi}\,\Gamma\left(\frac{n-1}{2}\right)\sigma_1^{n-1}} \cdot \frac{\mathrm{d}b_2}{\left\{\frac{\sigma_2^2}{\sigma_1^2}(1-\rho^2) + \left(b_2 - \frac{\rho\sigma_2}{\sigma_1}\right)^2\right\}^{n/2}}. \tag{102}$$

Demonstrație. Dacă în teorema III.13 facem transformarea

$$b_2 = r\frac{s_2}{s_1}$$

găsim probabilitatea elementară a statisticilor s_1, s_2, și b_2:

$$dF = C_1 \exp\left[-\frac{n}{2(1-\rho^2)}\left\{\frac{s_1^2}{\sigma_1^2} - \frac{2\rho s_1^2 b_2}{\sigma_1\sigma_2} + \frac{s_2^2}{\sigma_2^2}\right\}\right] s_1^{n-1} s_2^{n-3} \times$$

(103)

$$\times\left(1 - \frac{s_1^2 b_2^2}{s_2^2}\right)^{(n-4)/2} ds_1 ds_2 dr.$$

Integrînd (103) în raport cu s_2, obținem pentru s_1 și b_2 probabilitatea elementară:

$$dF = C_2 \exp\left[-\frac{n}{2(1-\rho^2)}\left\{\frac{s_1^2}{\sigma_1^2} - \frac{2\rho^2 s_1^2 b_2}{\sigma_1\sigma_2} + \right.\right.$$

(104)

$$\left.\left. + \frac{s_1^2 b_2^2}{\sigma_2^2}\right\}\right] s_1^{n-1} ds_1 db_2.$$

Integrînd (104) în raport cu s_1, obținem pentru b_2 probabilitatea elementară

$$dF = C_3 \frac{db_2}{\left(1 - \frac{2\rho\sigma_1}{\sigma_2} b_2 + \frac{\sigma_1^2}{\sigma_2^2} \cdot b_2^2\right)^{n/2}}.$$

Determinînd constantele C_1, C_2 și C_3, teorema este demonstrată.

Observație. (102) este simetrică față de punctul $\beta_2 = \frac{\rho\sigma_2}{\sigma_1}$, coeficientul de regresie al lui Y asupra lui X.

TEOREMA 8. *În selecții din populații normale, statistica*

(105) $$t = u\sqrt{n-2},$$

unde

(106) $$u = \frac{(b_2 - \beta_2)s_1}{(s_2^2 - b_2^2 s_1^2)^{1/2}}$$

este repartizată $t(n-2)$.

Demonstrație. Din teorema III.13, avem că statisticile a, b și c definite în (III.14) au probabilitatea elementară (III.15)

(107) $$dF \simeq \exp\left[-a + 2\rho b - c\right](ac - b^2)^{\frac{1}{2}(n-4)} da db dc.$$

Din (106) avem

$$u = \frac{b - \rho a}{\sqrt{(ac - b^2)}}.$$

Scanned with OKEN Scanner

Dacă în (107) exprimăm pe c în funcţie de u, avem :

$$dF = \frac{\exp\left[-a\left(1+\frac{\rho^2}{u^2}\right)\right] da\,du}{a u^{n-1}} \cdot \exp\left[\frac{u^2+1}{u^2}\left(2\rho b - \frac{b^2}{a}\right)\right](b-\rho a)^{n-2}\,db. \tag{108}$$

Integrînd a doua parte a membrului drept al relaţiei (108) găsim un factor proporţional cu

$$\exp\left\{\frac{a\rho^2(u^2+1)}{u^2}\right\}\left(\frac{au^2}{u^2+1}\right)^{(n-1)/2}. \tag{109}$$

Din (108) şi (109) găsim probabilitatea elementară a statisticilor a şi u.

$$dF \simeq \frac{a^{(n-1)/2}\exp(-a+\rho^2 a)\,da}{a} \cdot \frac{du}{(1+u^2)^{(n-1)/2}}. \tag{110}$$

Deci a şi u sînt independente şi pentru u avem

$$dF \simeq \frac{du}{(1+u^2)^{(n-1)/2}}, \tag{111}$$

ceea ce demonstrează teorema.

§ 6. Coeficienţii de selecţie

Să notăm prin r, b şi s^2 coeficienţii de selecţie corespunzători coeficienţilor ρ, β şi respectiv σ^2.

Coeficienţii b sînt determinaţi prin minimizarea expresiei

$$ns^2_{1.23\ldots p} = \sum_{i=1}^{n}\left(x_{1i} - \sum_{j=2}^{p} b_{ij\cdot q_j} x_{ji}\right)^2 \tag{112}$$

şi ca în (71) avem :

$$r^2_{ij.k} = b_{ij\cdot k} b_{ji\cdot k}. \tag{71'}$$

Analog cu (90), (91) şi (92) avem

$$\frac{s^2_{1.2\ldots p}}{s^2_1} = (1-r^2_{12})(1-r^2_{13.2})(1-r^2_{14.23})\ldots(1-r^2_{1p.23\ldots(p-1)}). \tag{90'}$$

$$b_{12\cdot pr} = \frac{b_{12\cdot r} - b_{1p\cdot r} b_{p2\cdot r}}{1 - b_{2p\cdot r} b_{p2\cdot r}}, \tag{91'}$$

Scanned with OKEN Scanner

și respectiv

$$r_{12\cdot pr} = (b_{12\cdot pr}b_{21\cdot pr})^{1/2} = \frac{r_{12\cdot r} - r_{1p\cdot r}r_{2p\cdot r}}{\{(1 - r^2_{1p\cdot r})(1 - r^2_{2p\cdot r})\}^{1/2}}. \tag{92'}$$

Dacă egalăm cu zero derivatele expresiei (112) în raport cu b_{1j}, avem următoarele $(p - 1)$ ecuații

$$\sum_{i=1}^{n} x_{ji}\left(x_{1i} - \sum_{j=2}^{p} b_{1j\cdot q_j}x_{ji}\right) = 0,\ j = 2, 3, \ldots, p$$

pe care le mai putem scrie

$$\sum x_j x_{1\cdot 23\ldots p} = 0,\ j = 2, 3, \ldots, p \tag{113}$$

unde $x_{1\cdot 23\ldots p}$ este rezidul de la regresia ajustată.
Din (113) urmează că

$$\sum x^2_{1\cdot q} = \sum x_{1\cdot q}(x_1 - \sum b_{ij\cdot q_j}x_j) = \sum x_{1\cdot q}x_1 \tag{114}$$

și asemănător

$$\sum x_{1\cdot r}x_{2\cdot r} = \sum x_{1\cdot r}x_2 = \sum x_1 x_{2\cdot r}, \tag{115}$$

unde r este orice mulțime de indici.

Considerăm repartițiile corelațiilor parțiale de selecție și a coeficienților de regresie de selecție în cazul normal.

PROPOZIȚIA 16. *Repartiția comună a oricăror două erori $x_{1\cdot m}$ și $x_{2\cdot m}$, pentru orice mulțime de indici m, este normală bidimensională cu coeficientul de corelație $\rho_{12\cdot m}$.*

Din (XXVIII.64), (101) și din propoziția 16 obținem :

COROLARUL 1. *Avem :*

$$D^2(r_{12\cdot m}) = \frac{1}{n}(1 - \rho^2_{12\cdot m})^2 \tag{116}$$

$$D^2(b_{12\cdot m}) = \frac{1}{n}\frac{\sigma^2_{1\cdot m}}{\sigma^2_{2\cdot m}}(1 - \rho^2_{12\cdot m}) = \frac{1}{n}\frac{\sigma^2_{1\cdot 2m}}{\sigma^2_{2\cdot m}}. \tag{117}$$

PROPOZIȚIA 17. *Statistica*

$$t = (b_{12\cdot m} - \beta_{12\cdot m})\left\{\frac{s^2_{2\cdot m}(n - d - 2)}{s^2_{1\cdot m} - b^2_{12\cdot m}s^2_{2\cdot m}}\right\}^{1/2} \tag{118}$$

unde d este numărul de indici din mulțimea m, are o repartiție $t(n - d - 2)$.

Scanned with OKEN Scanner

§ 9. Teste pentru rapoartele de corelație și liniaritatea regresiei

În § 4 am văzut că $\eta_1^2 = \rho^2$ indică că curba de regresie nu este mai bună decît linia dreaptă și deci că $\eta_1^2 - \rho^2$ este un indicator al neliniarității regresiei. Deci pentru un test privind liniaritatea regresiei lui X asupra lui Y, putem folosi statistica $e_1^2 - r^2$, e_1^2 fiind raportul de corelație observat al variabilei X asupra lui Y și definit de

$$e_1^2 = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2} = \frac{\sum n_i \bar{x}_i^2 - n\bar{x}^2}{\sum\sum x_{ij}^2 - n\bar{x}^2}. \tag{119}$$

PROPOZIȚIA 18. *Avem*

$$0 \leqslant r^2 \leqslant e_i^2 \leqslant 1, \quad i = 1, 2. \tag{120}$$

În cele ce urmează vom da un test pentru verificarea ipotezei $\eta_1^2 = 0$ și de asemenea îl vom pune în relație cu testul pentru verificarea ipotezei $\rho = 0$ dat în XXIX § 2. Aceste probleme au fost rezolvate mai întîi de Fisher.

După (120), identitatea

$$ns_1^2 \equiv ns_1^2 r^2 + ns_1^2(e_1^2 - r^2) + ns_1^2(1 - e_1^2) \tag{121}$$

are toți termenii din dreapta ne-negativi. Deoarece

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\{(\bar{x}_i - \bar{x}) - b_i(\bar{y}_i - \bar{y})\}^2 \equiv \sum_i\sum_j (x_i - \bar{x})^2 - r^2\sum_i\sum_j (x_{ij} - \bar{x})^2,$$

(121) poate fi scris în x ca

$$\sum_i\sum_j (x_{ij} - \bar{x})^2 \equiv r^2\sum_i\sum_j (x_{ij} - \bar{x})^2 + \sum_i\sum_j\{(\bar{x}_i - \bar{x}) - \\ - b_1(\bar{y}_i - \bar{y})\}^2 + \sum_i\sum_j (x_{ij} - \bar{x}_i)^2. \tag{122}$$

Am descompus o formă pătratică în x_{ij} în alte trei astfel de forme pătratice. Acum vom presupune că y_i sînt fixate și că toți x_{ij} sînt independenți și repartizați normali cu aceeași dispersie (care fără a pierde din generalitate o presupunem 1). Pentru moment nu punem problema mediilor lui x_{ij}.

În ipoteza H_0 conform căreia fiecare x_{ij} are aceeași medie, adică curba de regresie este o linie paralelă cu axa x-lor,

$$\sum\sum (x_{ij} - \bar{x})^2 \sim \chi^2(n - 1). \tag{123}$$

Scanned with OKEN Scanner

PROPOZIȚIA 19. *Formele pătratice din membrul drept al relației* (122) *sînt independente și repartizate* χ^2 cu 1, $k-2$ și *respectiv* $n-k$ *grade de libertate.*

Demonstrația urmează din teorema lui Cochran (teorma IV.20), (123) și din faptul că suma gradelor de libertate corespunzătoare acestor forme pătratice este $n-1$.

COROLARUL 2. *Avem*

$$\frac{r^2/1}{(1-r^2)/(n-2)} \sim F(1, n-2), \tag{124}$$

$$\frac{e_1^2/(k-1)}{(1-e_1^2)/(n-k)} \sim F(k-1, n-k). \tag{125}$$

Atît testul bazat pe (124) cît și testul bazat pe (125) se folosesc pentru verificarea ipotezei H_0, însă în special (124) pentru $\rho^2 = 0$, în timp ce (125) pentru $\gamma_1^2 = 0$.

Dacă ipoteza alternativă este ipoteza conform căreia regresia lui x asupra lui y este liniară testul bazat pe (124) va avea o putere mare, însă dacă ipoteza alternativă este orice altă ipoteză diferită de H_0, (125) este mai puternic.

Dacă variabilele x_{ij} nu au toate aceeași medie, membrul stîng al relației (122) nu mai este repartizat $\chi^2(n-1)$. Totuși dacă trecem primul termen din dreapta în stînga, obținem

$$ns_1^2(1-r^2) \equiv ns_1^2(e_1^2 - r^2) + ns_1^2(1-e_1^2). \tag{126}$$

Deoarece

$$ns_1^2(1-r^2) \equiv \sum_i \sum_j \{x_{ij} - (a_1 + b_1 y_i)\}^2,$$

suma pătratelor rezidurilor de la regresia liniară ajustată, în ipoteza H_0' conform căreia regresia lui x asupra lui y este exact liniară,

$$ns_1^2(1-r^2) \sim \chi^2(n-2), \tag{127}$$

fiind pierdut cîte un grad de libertate pentru fiecare parametru ajustat în dreapta de regresie.

PROPOZIȚIA 20. *Formele pătratice din membrul drept al relației* (126) *sînt independete și repartizate* χ^2 *cu* $(k-2)$, *respectiv* $n-k$ *grade de libertate.*

Demonstrația rezultă din teorema IV.20, ținînd seama de (127) și de faptul că rangurile formelor pătratice din membrul drept al relației (126) sînt $(k-2)$ și respectiv $(n-k)$.

COROLARUL 3. *Statistica*

$$\frac{(e_1^2 - r^2)/(k-2)}{(1-e_1^2)/(n-k)} \tag{128}$$

este repartizată $F(k-2, n-k)$.

Scanned with OKEN Scanner

(128) poate fi folosită ca bază a unui test pentru verificarea ipotezei H_0' conform căreia regresia este liniară, H_0' fiind respinsă pentru valori mari ale statisticii test. Astfel intuiția noastră că $e_1^2 - r^2$ trebuie să ofere un test pentru liniaritatea regresiei este corectă, însă (128) arată că testul rezultat va fi o funcție de $(1 - e_1^2)$, k și n, astfel că valoarea lui $(e_1^2 - r^2)$ singură reprezintă puțin.

Observație. Toate cele trei teste prezentate sînt teste ale raportului de verosimilități. De exemplu ipoteza conform căreia toate variabilele x_{ij} au aceeași medie poate fi privită în două moduri:

— putem privi variabilele ca variind pe o linie dreaptă care are doi parametri, și verificăm ipoteza conform căreia linia are panta zero, care impune o constrîngere asupra celor doi parametri. Avem $k = 2$ și $r = 1$ și astfel obținem un $F(1, n - 2)$; aceasta este (124);

— cele k medii condiționate sînt pe o curbă dată de un polinom de grad $(k - 1)$ și verificăm ipoteza conform căreia toți coeficienții, exceptînd constanta, sînt zero; aceasta implică $(k - 1)$ constrîngeri; aceasta dă (125).

— în fine dacă în a doua formulare verificăm ipoteza conform căreia toți coeficienții polinomului, exceptînd termenul constant și cel liniar, sînt zero, astfel încît mediile condiționate variază pe o linie dreaptă, impunei $(k - 2)$ constrîngeri și obținem un $F(k - 2, n - k)$; acesta este (128).

Urmează că pentru y_i fixați, aceste teste sînt U.C.M.P.

§ 10. Aplicații ale proprietății de nedeplasare la regresie

Presupunem că pentru un x fixat, Y este repartizată normal cu media o funcție de x, regresie lui Y asupra lui x și care este presupusă a fi liniară

$$M(Y \mid x) = \alpha + \beta x$$

și cu dispersia σ^2 cunoscută.

Punînd $v_i = (x_i - \bar{x})/\sqrt{\sum (x_j - \bar{x})^2}$ și $\gamma + \delta v_i = \alpha + \beta x_i$, astfel încît $\sum v_i = 0$, $\sum v_i^2 = 1$, și

$$\alpha = \gamma - \delta \frac{\bar{x}}{\sqrt{\sum (x_j - \bar{x})^2}}, \quad \beta = \frac{\delta}{\sqrt{\sum (x_j - \bar{x})^2}}$$

densitatea de repartiție comună a variabilelor $Y_1, \ldots, Y_n$ este

$$\frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left[-\frac{1}{2\sigma^2} \sum (y_i - \gamma - \delta v_i)^2 \right]. \tag{129}$$

Aceste densități constituie o familie exponențială (XVI.1) cu

$$U = \sum v_i Y_i, \quad T_1 = \sum Y_i^2, \quad T_2 = \Sigma Y_i$$

$$\theta = \delta/\sigma^2, v_1 = -1/2\sigma^2, \quad v_2 = \gamma/\sigma^2. \tag{130}$$

Scanned with OKEN Scanner

TEOREMA 9. *Există teste* U.C.M.P. *nedeplasate pentru verificarea ipotezei*

$$a\gamma + b\delta = c$$

unde a, b *şi* c *sînt constante date şi prin urmare a intervalelor de încredere pentru parametrul*

$$\rho = a\gamma + b\delta.$$

Demonstraţia rezultă din (129), (130) şi din teorema XVI.1.

Pentru a obţine explicit aceste intervale de încredere, avem nevoie de testul U.C.M.P. nedeplasat pentru verificarea ipotezei

$$H_0 : \rho = \rho_0$$

care este dat de regiunea de acceptare

$$\frac{|b\sum v_i Y_i + a\bar{Y} - \rho_0|/\sqrt{(a^2/n) + b^2}}{\sqrt{[\sum (Y_i - \bar{Y})^2 - (\sum v_i Y_i)^2]/(n-2)}} \leqslant C$$

unde

$$\int_{-C}^{C} g_{n-1}(t)dt = 1 - \alpha,$$

g_{n-1} fiind densitatea unei variabile $t(n-2)$.

Intervalele de încredere pentru ρ rezultate sînt centrate în $b\sum v_i Y_i + a\bar{Y}$ şi au lungimea

$$L = 2C\sqrt{[(a^2/n) + b^2][\sum (Y_i - \bar{Y})^2 - (\sum v_i Y_i)^2]/(n-2)}. \tag{132}$$

Cum

$$[\sum (Y_i - \bar{Y})^2 - (\sum v_i Y_i)^2]/\sigma^2 \sim \chi^2(n-2).$$

urmează că

$$M(L) = 2C_n \sigma \sqrt{(a^2/n) + b^2}). \tag{133}$$

În aplicaţii practice, a şi b sînt funcţii de x. Dacă acestea sînt la dispoziţia experimentatorului şi prin urmare există o alegere în raport cu a şi b, $M(L)$ se minimizează, minimizînd $(a^2/n) + b^2$. Vom considera $(a^2/n) + b^2$ ca inversa măsurii acurateţii unui interval.

Exemplul 10. Intervalele de încredere pentru panta $\beta = \delta/\sqrt{\sum (x_j - \bar{x})^2}$ sînt obţinute din intervalele anterioare luînd

$$a = 0 \text{ şi } b = 1/\sqrt{\sum (x_j - \bar{x})^2}.$$

În acest caz acurateţea creşte cu $\sum (x_j - \bar{x})^2$ şi dacă x_i trebue să fie ales dintr-un interval $[C_0, C_1]$, $M(L)$ este minimizat punînd cîte o jumătate din observaţii în fiecare extremitate.

Scanned with OKEN Scanner

Exemplul 11. Un alt parametru de interes este valoarea $\alpha + \beta x_0$ în $x = x_0$. Deoarece

$$\alpha + \beta x_0 = \gamma + \delta(x_0 - \bar{x})/\sqrt{\sum(x_j - \bar{x})^2},$$

constantele a şi b sînt $a = 1$, $b = (x - x_0)/\sqrt{\sum(x_j - x_0)^2}$. Acurateţea maximă se obţine minimizînd $|\bar{x} - x_0|$.

Exemplul 12. Obişnuit prezintă interes estimarea punctului x în care $\alpha + \beta x$ are o anumită valoare dată. De exemplu dorim să estimăm dozajul $x = -\alpha/\beta$ în care $M(Y|x) = 0$, sau echivalent valoarea $v = (x - \bar{x})/\sqrt{\sum(x_j - \bar{x})^2}$ în care $\gamma + \delta v = 0$. Intervalul cu cea mai mare acurateţe pentru soluţia $-\gamma/\delta$ a acestei ecuaţii poate fi obţinut din testul U.C.M.P. nedeplasat pentu ipoteza $-\gamma/\delta = v_0$. Regiunile de acceptare ale acestor teste sînt date de (131) cu $a = 1$, $b = v_0$ şi $\rho_0 = 0$ şi mulţimile de încredere rezultate pentru v sînt mulţimile de valori v ce satisfac

$$v^2[C^2S^2 - (\sum v_i Y_i)^2] - 2v\bar{Y}(\sum v_i Y_i) - \frac{1}{n}(C^2S^2 + n\bar{Y}^2) \geqslant 0$$

unde $S^2 = [\sum(Y_i - \bar{Y})^2 - (\sum v_i Y_i)^2]/(n-2)$. Dacă ecuaţia pătratică în v asociată are rădăcinile $\underline{v}$, $\bar{v}$

$$\underline{v} \leqslant v \leqslant \bar{v} \text{ cînd } |\sum v_i Y_i|/S > C$$

şi

$$v \leqslant \underline{v} \text{ sau } v \geqslant \bar{v} \text{ cînd } |\sum v_i Y_i|/S < C.$$

Cînd dreapta $y = \gamma + \delta v$ este aproape paralelă cu axa v, intercepţia cu axa v va fi mare în valoare absolută, însă semnul său poate fi schimbat printr-o schimbare mică a unghiului. Aceasta este o posibilitate ca discriminantul polinomului pătratic să fie negativ

$$n\bar{Y}^2 + (\sum v_i Y_i)^2 < C^2S^2,$$

în care caz ecuaţia pătratică asociată nu are soluţii. În acest caz mulţimea de încredere devine întreaga axă.

§ 11. Modele ale regresiei pentru date direcţionale

Forma de dependenţă ce apare cînd $\rho^2 = c$ (descrisă în teorema XXVIII.17. (V)) sugerează folosirea modelelor regresionale de forma:

$$t_2(\hat{Y}) = At_1(X) + b \tag{134}$$

pentru variabilele aleatoare X şi Y pe subspaţiile generale M şi N. A şi b sînt supuse la constrîngerea ca $\hat{Y}$, valoarea predicată a lui Y, să varieze în N pentru toţi X în M. Un mod natural de a alege A şi b este metoda C.M.M.P. în spaţiul lui t_2.

Scanned with OKEN Scanner

Cazul circular-liniar. Pentru o variabilă unghiular θ și o variabilă liniară x, obținem

$$\hat{x} = a_1 \cos\theta + a_2 \sin\theta + b \tag{135}$$

și nici o restricție asupra lui a_1, a_2, și b (Mardia (1976)). Estimațiile de C.M.M.P. obținute din ecuațiile normale obișnuite sînt

$$\hat{a}_1 = \frac{\sigma}{V_c} \frac{(\rho_{cx} - \rho_{cs}\rho_{sx})}{(1 - \rho_{cs}^2)}, \quad \hat{a}_2 = \frac{\sigma}{V_s} \cdot \frac{(\rho_{sx} - \rho_{sc}\rho_{cx})}{(1 - \rho_{cs}^2)},$$

$$\hat{b} = \bar{x} - \bar{R}(\hat{a}_1 \cos\bar{x}_0 + \hat{a}_2 \sin\bar{x}_0),$$

unde $D^2(x) = \sigma^2$, $D^2(\cos\theta) = V_c$, coef. corel. dintre $\cos\theta$ și $x = \rho_{cx}$, ... Pentru modelul regresiei în care x_i sînt independente și repartizate $N(a_1 \cos\theta_i + a_2 \sin\theta_i + b, \sigma^2)$, aceste estimații de C.M.M.P. sînt de asemenea estimații de V.M.

Observație. Modelele cilindrice (Mardia, Sutton (1978)) au funcțiile de regresie de forma (135).

Cazurile circular-circular și sferic-sferic

Pentru variabilele unghiulare θ și ϕ avem

$$t_1'(\theta) = (\cos\theta, \sin\theta), \ldots!$$

și (134) se reduce la

$$\hat{\phi} = \pm\theta - \alpha \text{ sau } \hat{\phi} = \text{const.}$$

Mai în general, pentru x și y vectori aleatori pe sfera unitate din R^3 sau hipersfera din R^p, cu $t_1(x) = x$, $t_2(y) = y$, (134) revine la $A = 0$ sau $AA' = I_p$ astfel că pentru regresia care nu-i constantă $\hat{y} = Ax$ cu A ortogonală. A lua estimația de C.M.M.P. pentru A este echivalent cu a alege matricea ortogonală A să maximizeze

$$M(y'Ax) = \text{Ur}\{AM(xy')\}.$$

Astfel o soluție este

$\hat{A} = (B'B)^{-1/2}B'$ unde $B = M(xy')$ (Downs(1972)).

Dacă rang $\{M(xy')\} = p$, această soluție este unică.

Pentru modelul regresiei în care y_i sînt independente și cu repartiție von Mises $M(Ax_i, \varkappa)$ atunci estimațiile de C.M.M.P. sînt de asemenea estimații de C.M.M.P. Dacă $\varkappa_2 = 0$, repartițiile bidimensionale von Mises au funcțiile de regresie de forma $\hat{y} = Ax$.

Scanned with OKEN Scanner

Capitolul XXXI

REGRESIE LINIARĂ. MODELUL DE RANG MAXIM

§ 1. Definiţia modelului liniar general. Presupuneri

Considerăm modelul

$$(1) \qquad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \text{ pentru } i = 1, 2, \ldots, n.$$

Facem următoarea schimbare în notaţii: în locul lui β_0 vom scrie $\beta_0 X_{i0}$, toate valorile lui X_{i0} fiind egale cu unu. Aceasta dă

$$(1') \qquad Y_i = \beta_0 X_{i0} + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

pentru $i = 1, 2, \ldots, n$, cu $X_{i0} = 1$ pentru toţi i.

Definim

$$(2) \quad \mathbf{X} = \begin{pmatrix} X_{10} & X_{11} & X_{12} \\ X_{20} & X_{21} & X_{22} \\ \cdots & \cdots & \cdots \\ X_{n0} & X_{n1} & X_{n2} \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \text{ şi } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Atunci sistemul reprezentat de (1') este

$$(3) \qquad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

În cazul a k variabile independente

$$(4) \qquad \mathbf{X} = \begin{pmatrix} X_{10} & X_{11} & \ldots & X_{1k} \\ X_{20} & X_{21} & \ldots & X_{2k} \\ \vdots & \vdots & & \vdots \\ X_{n0} & X_{n1} & \ldots & X_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix},$$

$\mathbf{Y}$ şi $\boldsymbol{\varepsilon}$ fiind definiţi ca în (2). Ecuaţia (3) este de asemenea neschimbată şi reprezintă modelul indiferent de numărul variabilelor $\mathbf{X}$ atîta timp cît $k < n$.

Scanned with OKEN Scanner

DEFINIȚIA 1. Modelul stocastic

$$Y = \sum_{i=0}^{k} \beta_i X_i + \varepsilon \tag{5}$$

unde: $\{X_i : i = 0, 1, 2, \ldots, k\}$ este o mulțime de variabile nealeatoare (numite variabile explicative sau regresoare sau independente),
ε este o variabilă aleatoare neobservată și
$\{\beta_i : i = 0, 1, 2, \ldots, k\}$ sînt constante însă fixe,
poartă numele de modelul liniar general.

DEFINIȚIA 2. β_i, $i = 0, 1, \ldots, k$ poartă numele de parametrii modelului (sau de coeficienții de regresie), iar Y de variabilă dependentă sau variabilă ce trebuie explicată.

DEFINIȚIA 3. Vom spune că modelul este liniar sau neliniar, după cum este liniar sau neliniar în parametrii.

DEFINIȚIA 4. Vom numi ordinul modelului, valoarea celei mai mari puteri a variabilei independente.

Presupunerile în care lucrăm sînt:

$$M(\varepsilon) = 0 \tag{6}$$

$$\mathscr{C}(\varepsilon) = M[\varepsilon - M(\varepsilon)]\,[\varepsilon - M(\varepsilon)]' = M(\varepsilon\varepsilon') = \sigma^2 \mathbf{I}_n.$$

Deoarece $M(\varepsilon) = 0$ un alt mod de a scrie modelul (3) este

$$M(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}. \tag{3'}$$

§ 2. Estimarea parametrilor

2.1. Metoda C.M.M.P.

Suma pătratelor abaterilor observațiilor de la valorile lor medii este:

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = [\mathbf{Y} - M(\mathbf{Y})]'\,[\mathbf{Y} - M(\mathbf{Y})] = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\,(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) =$$

$$= \mathbf{Y}'\mathbf{Y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

și cum $\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y}$ este o matrice 1×1, $(\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y})' = \mathbf{Y}'\mathbf{X}\boldsymbol{\beta}$ și

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \mathbf{Y}'\mathbf{Y} - 2\,\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}. \tag{7}$$

Estimația de C.M.M.P. a parametrului $\boldsymbol{\beta}$ este valoarea $\mathbf{b}$ care cînd este înlocuită în (7), face minimă $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$. $\mathbf{b}$ poate fi determinată derivînd (7) în raport cu $\boldsymbol{\beta}$ și rezolvînd ecuația matricială obținută egalînd cu 0 ecuația rezultată (și în același timp înlocuind $\boldsymbol{\beta}$ prin $\mathbf{b}$). A deriva $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ în raport cu vectorul $\boldsymbol{\beta}$ este echivalent cu a deriva $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ separat în raport cu fiecare ele-

Scanned with OKEN Scanner

ment al vectorului β și rearanjînd totul în formă matricială. Aceasta conduce la *ecuațiile normale*:

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{Y}. \tag{8}$$

Dacă $(\mathbf{X}'\mathbf{X})^{-1}$ există, atunci (8) are o soluție unică și anume

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{9}$$

PROPOZIȚIA 1. *Estimatorul de* C.M.M.P. *al parametrului* β *din modelul* (3) *este vectorul* $\mathbf{b}$ *ce minimizează* (7) *și este dat de* (9).

În acest caz, spunem că avem *modelul de rang maxim.* Cînd $\mathbf{X}'\mathbf{X}$ este de rang maxim, soluția ecuațiilor (8) poate fi scrisă ca (9). Pe de altă parte, dacă $(\mathbf{X}'\mathbf{X})^{-1}$ nu există, soluția (8) poate fi scrisă cu ajutorul inversei generalizate a matricei $\mathbf{X}'\mathbf{X}$. Acesta este cazul *modelului care nu este de rang maxim.*

Din (4), $\mathbf{X}'\mathbf{X}$ este o matrice pătrată de ordin $k+1$, avînd ca elemente sume de produse, sumele fiind după i de la 1 pînă la n.

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \sum X_{i0}^2 & \sum X_{i0}X_{i1} & \dots & \sum X_{i0}X_{ik} \\ \sum X_{i0}X_{i1} & \sum X_{i1}^2 & \dots & \sum X_{i1}X_{ik} \\ \vdots & \vdots & & \vdots \\ \sum X_{i0}X_{ik} & \sum X_{i1}X_{ik} & \dots & \sum X_{ik}^2 \end{pmatrix} \tag{10}$$

și

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum X_{i0}Y_i \\ \sum X_{i1}Y_i \\ \vdots \\ \sum X_{ik}Y_i \end{pmatrix} \tag{11}$$

Deoarece $X_{i0} = 1$ pentru toți $i = 1, 2, \dots, n$ și deoarece toate sumările se fac de la 1 pînă la n, $\sum X_{i0}^2 = n$, $\sum X_{i0}X_{i1} = X_{\cdot 1}$ și $\sum X_{i0}Y_i = Y_\cdot$,

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} N & X_{\cdot 1} & X_{\cdot 2} & \dots & X_{\cdot k} \\ X_{\cdot 1} & \sum X_{i1}^2 & \sum X_{i1}X_{i2} & \dots & \sum X_{i1}X_{ik} \\ X_{\cdot 2} & \sum X_{i1}X_{i2} & \sum X_{i2}^2 & \dots & \sum X_{i2}X_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ X_{\cdot k} & \sum X_{i1}X_{ik} & \sum X_{i2}X_{ik} & \dots & \sum X_{ik}^2 \end{pmatrix} \tag{12}$$

DEFINIȚIA 5. Modelul în care toți X — și sînt zero poartă numele de model cu intercepție; intercepția este β_0 valoarea lui $M(Y)$.

DEFINIȚIA 6. Modelul în care nu apare termenul β_0 poartă numele de model fără intercepție.

Scanned with OKEN Scanner

Dacă în (4) punem $X_{i0} = 1$ pentru toţi i, prima coloană a matricei $\mathbf{X}$ va fi formată numai din 1. Definind

$$\mathbf{1}_n = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad \text{şi} \quad \mathbf{X}_1 = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{pmatrix} \tag{13}$$

$\mathbf{X}$ poate fi scris ca

$$\mathbf{X} = (\mathbf{1} \quad \mathbf{X}_1), \tag{14}$$

unde $\mathbf{1}$ este de ordin n şi $\mathbf{X}_1$ este o matrice $n \times k$.

Cu notaţia

$$\bar{\mathbf{X}}' = (\bar{X}_{\cdot 1}, \bar{X}_{\cdot 2}, \dots, \bar{X}_{\cdot k}), \tag{15}$$

$$\mathbf{1}_n'\mathbf{1} = n, \quad \mathbf{1}'\mathbf{Y} = n\bar{Y} \quad \text{şi} \quad \mathbf{1}'\mathbf{X}_1 = n\bar{\mathbf{X}}' \tag{16}$$

unde am renotat media $\bar{Y}_{\cdot}$ prin $\bar{Y}$.

Atunci $\mathbf{b}$ poate fi exprimat ca

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \left(\begin{pmatrix} \mathbf{1}' \\ \mathbf{X}_1' \end{pmatrix}(\mathbf{1}' \quad \mathbf{X}_1')\right)^{-1}\begin{pmatrix} \mathbf{1}' \\ \mathbf{X}_1' \end{pmatrix}\mathbf{Y} =$$

$$= \begin{pmatrix} n & \bar{\mathbf{X}}' \\ n\bar{\mathbf{X}}' & \mathbf{X}_1'\mathbf{X}_1 \end{pmatrix}^{-1}\begin{pmatrix} n\bar{Y} \\ \mathbf{X}_1'\mathbf{Y} \end{pmatrix}. \tag{17}$$

Cu

$$\mathbf{S} = \mathbf{X}_1'\mathbf{X}_1 - n\bar{\mathbf{X}}\bar{\mathbf{X}}' \tag{18}$$

obţinem

$$\mathbf{b} = \begin{pmatrix} 1/n + \bar{\mathbf{X}}'\mathbf{S}^{-1}\bar{\mathbf{X}} & -\bar{\mathbf{X}}'\mathbf{S}^{-1} \\ -\mathbf{S}^{-1}\bar{\mathbf{X}} & \mathbf{S}^{-1} \end{pmatrix}\begin{pmatrix} n\bar{Y} \\ \mathbf{X}_1'\mathbf{Y} \end{pmatrix}. \tag{19}$$

Considerînd partiţia:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}^* \end{pmatrix}, \tag{2'}$$

Scanned with OKEN Scanner

(19) poate fi scrisă ca

$$\begin{pmatrix} b_0 \\ \mathbf{b}^* \end{pmatrix} = \left(\begin{pmatrix} 1/n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} -\bar{\mathbf{X}}' \\ \mathbf{I} \end{pmatrix} \mathbf{S}^{-1} (-\bar{\mathbf{X}} \quad \mathbf{I}) \right) \begin{pmatrix} n\bar{Y} \\ \mathbf{X}_1'\mathbf{Y} \end{pmatrix} =$$

(19′)

$$= \begin{pmatrix} \bar{Y} - \bar{\mathbf{X}}'\mathbf{S}^{-1}(\mathbf{X}_1'\mathbf{Y} - n\bar{Y}\bar{\mathbf{X}}) \\ \mathbf{S}^{-1}(\mathbf{X}_1'\mathbf{Y} - n\bar{Y}\bar{\mathbf{X}}) \end{pmatrix}.$$

Urmează că

(21) $$\mathbf{b}^* = \mathbf{S}^{-1}(\mathbf{X}_1'\mathbf{Y} - n\bar{Y}\bar{\mathbf{X}})$$

și

(22) $$b_0 = \bar{Y} - \bar{\mathbf{X}}'\mathbf{b}^*.$$

Avem

(23) $$\mathbf{X}_1'\mathbf{X}_1 = \begin{pmatrix} \sum X_{i1}^2 & \sum X_{i1}X_{i2} & \dots & \sum X_{i1}X_{ik} \\ \sum X_{i1}X_{i2} & \sum X_{i2}^2 & \dots & \sum X_{i2}X_{ik} \\ \vdots & \vdots & & \vdots \\ \sum X_{i1}X_{ik} & \sum X_{i2}X_{ik} & \dots & \sum X_{ik}^2 \end{pmatrix},$$

$$n\bar{\mathbf{X}}\bar{\mathbf{X}}' = \{n\bar{X}_{\cdot p}\bar{X}_{\cdot q}\} \quad \text{pentru} \quad p, q = 1, 2, \dots, k,$$

ceea ce permite să scriem

$$\mathbf{S} = \{\sum_i X_{ip}X_{iq} - n\bar{X}_{\cdot p}\bar{X}_{\cdot q}\} \quad \text{pentru} \quad p, q = 1, 2, \dots, k.$$

Dacă definim :

(24) $$\mathscr{A} = \mathbf{X}_1 - 1_n\bar{\mathbf{X}}',$$

atunci

(25) $$\mathbf{S} = \mathscr{A}'\mathscr{A}$$

și

$$\mathbf{X}_1'\mathbf{Y} - n\bar{Y}\bar{\mathbf{X}} = \{\sum_i X_{ip}Y_i - n\bar{X}_{\cdot p}\bar{Y}\} \text{ pentru } p = 1, 2, \dots, k$$

$$= \mathscr{A}'\mathbf{Y},$$

ceea ce ne permite să scriem

(21′) $$\mathbf{b}^* = (\mathscr{A}'\mathscr{A})^{-1}\mathscr{A}'\mathbf{Y}.$$

Astfel ca în (22)

(22′) $$b_0 = \bar{Y} - \mathscr{A}'\bar{\mathbf{X}}.$$

Scanned with OKEN Scanner

2.2. Proprietățile estimatorului de C.M.M.P. pentru β

TEOREMA 1. i) *Elementele lui* $\mathbf{b}$ *sînt estimații nedeplasate pentru elementele lui* $\boldsymbol{\beta}$.

ii) $$\mathscr{C}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2.$$

Pentru demonstrație a se vedea teorema X.1.
Folosind partiția lui $(\mathbf{X}'\mathbf{X})^{-1}$ dată în (19) cu $\mathbf{S} = \mathscr{A}'\mathscr{A}$, ii) devine

$$\mathscr{C}\begin{pmatrix} b_0 \\ \mathbf{b}^* \end{pmatrix} = \begin{pmatrix} 1/n + \bar{\mathbf{X}}'(\mathscr{A}'\mathscr{A})^{-1}\bar{\mathbf{X}} & -\bar{\mathbf{X}}'(\mathscr{A}'\mathscr{A})^{-1} \\ -(\mathscr{A}'\mathscr{A})^{-1}\bar{\mathbf{X}} & (\mathscr{A}'\mathscr{A})^{-1} \end{pmatrix}\sigma^2.$$

Deci are loc
COROLARUL 1. *Statisticile* b_0 *și* $\mathbf{b}^*$ *au caracteristicile:*

(26) $$\mathscr{C}(\mathbf{b}^*) = (\mathscr{A}'\mathscr{A})^{-1}\sigma^2,$$

(27) $$D^2(b_0) = [1/n + \bar{\mathbf{X}}'(\mathscr{A}'\mathscr{A})^{-1}\bar{\mathbf{X}}]\sigma^2 = \sigma^2/n + \bar{\mathbf{X}}'\mathscr{C}(\mathbf{b}^*)\bar{\mathbf{X}}$$

și

(28) $$\operatorname{cov}(b_0, \mathbf{b}^{*\prime}) = -\bar{\mathbf{X}}'\mathscr{C}(\mathbf{b}^*).$$

Estimatorul $\mathbf{b}$ poate fi folosit pentru a estima $M(Y)$. Corespunzător modelului

$$M(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

avem

$$\widehat{M(Y)} = b_0 + b_1 X_1 + \ldots + b_k X_k.$$

Dacă

(29) $$\mathbf{X}_0' = (X_{00} \quad X_{01} \quad X_{02} \ldots X_{0k})$$

este o mulțime de valori X (cu $X_{00} = 1$) estimatorul pentru $M(Y)$ este

(30) $$\widehat{M(Y_0)} = b_0 + b_1 X_{01} + \ldots + b_k X_{0k} = \mathbf{X}_0'\mathbf{b}.$$

DEFINIȚIA 7. $\widehat{M(Y_0)}$ poartă numele de estimația valorii medii a lui Y corespunzătoare mulțimii $X_{00}, X_{01}, \ldots, X_{0k}$.
Corespunzător lui $M(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ din (3), ca un caz special al lui (30), avem

(30′) $$\widehat{M(\mathbf{Y})} = \mathbf{X}\mathbf{b}.$$

Acestea sînt estimațiile valorilor medii ale lui $\mathbf{Y}$ corespunzătoare la n valori observate. Adeseori aceste estimații poartă numele de valorile ajustate ale lui $\mathbf{Y}$ sau estimațiile valorilor $\mathbf{Y}$, (uneori în mod greșit deoarece (30) și (30′) sînt amîndouă estimații ale valorilor medii ale lui $\mathbf{Y}$).

Scanned with OKEN Scanner

TEOREMA 2. *Avem*

$$D^2[\widehat{M(Y_0)}] = \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\sigma^2 \tag{31}$$

şi

$$\mathscr{C}[\widehat{M(\mathbf{Y})}] = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2. \tag{32}$$

COROLARUL 2. *Avem*

$$\begin{aligned}\mathscr{C}[\widehat{M(\mathbf{Y})}] &= (\sigma^2/n)\,\mathbf{1}\mathbf{1}' + \mathscr{A}\mathscr{C}(\mathbf{b}^*)\,\mathscr{A}' \\ &= (\sigma^2/n)\,\mathbf{1}\mathbf{1}' + \mathscr{A}(\mathscr{A}'\mathscr{A})^{-1}\mathscr{A}'\sigma^2.\end{aligned} \tag{32}'$$

Demonstraţia rezultă ţinînd în (32) seama de (14) şi folosind (24). Corespunzător lui $\mathbf{X}_0'$, valoarea medie a lui $\mathbf{Y}$ este $M(Y_0)$, estimată de $\widehat{M(Y_0)}$ (dată de (30)). Dacă considerăm o altă observaţie Y_v (de exemplu) corespunzătoare vectorului $\mathbf{X}_v$, atunci după model, avem

$$Y_v = \mathbf{X}_v'\boldsymbol{\beta} + \varepsilon_v$$

unde ε_v este o variabilă aleatoare care nu poate fi nici observată, nici estimată. Deci cea mai bună predicţie pentru Y_v, pe care o vom nota prin $\widetilde{Y}_v$ este

$$\widetilde{Y}_v = \mathbf{X}_v'\mathbf{b}.$$

Astfel $\mathbf{X}_v'\mathbf{b}$ poate fi folosit ca predicţie a observaţiei următoare corespunzătoare lui $\mathbf{X}_v'$ cît şi ca estimator al lui $M(Y_v)$ corespunzătoare la $\mathbf{X}_v'$. Variaţia lui Y_v de la valoarea sa de predicţie este

$$Y_v - \widetilde{Y}_v = Y_v - \mathbf{X}_v'\mathbf{b} = \mathbf{X}_v'(\boldsymbol{\beta} - \mathbf{b}) + e_v.$$

Cum Y_v a fost obţinută ca o observaţie independentă de aceea folosită în obţinerea lui $\mathbf{b}$, $\mathbf{b}$ este independentă de e_j şi astfel

$$\operatorname{cov}(\mathbf{b}, e_j) = 0. \tag{33}$$

Deci

$$\begin{aligned}\mathscr{C}(Y_v - \widetilde{Y}_v) &= \mathbf{X}_v'\mathscr{C}(\boldsymbol{\beta} - \mathbf{b})\,\mathbf{X}_v + D^2(e_j) = \\ &= [\mathbf{X}_v'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_v + 1]\,\sigma^2.\end{aligned} \tag{34}$$

Astfel estimaţia valorii medii a lui $\mathbf{Y}$ corespunzătoare lui $\mathbf{X}_v$ este

$$\widehat{M(Y_v)} = \mathbf{X}_v'\mathbf{b} \text{ (ca în (30))} \tag{30''}$$

Scanned with OKEN Scanner

cu dispersia

$$D^2(\widehat{M(Y_v)}) = \mathbf{X}_v'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_v\sigma^2 \quad \text{(ca în (31))} \tag{32''}$$

și valoarea de predicție a unei observații corespunzătoare lui $\mathbf{X}_v$ are aceeași valoare, $\mathbf{X}_v'\mathbf{b} = \widehat{Y}_v$ cu dispersia abaterilor valorilor $\mathbf{Y}$ (corespunzătoare la $\mathbf{X}_v$) de la această predicție,

$$[\mathbf{X}_v'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_v + 1]\,\sigma^2.$$

Aceste rezultate sînt adevărate pentru orice valoare a lui $\mathbf{X}_v$.

2.3. Sume de pătrate. Estimarea lui σ^2

Dacă folosim convenția

$$\hat{\mathbf{Y}} \equiv \widehat{M(\mathbf{Y})} = \mathbf{X}\mathbf{b}, \tag{35}$$

atunci

$$\begin{aligned}\mathbf{Y} - \hat{\mathbf{Y}} &= \mathbf{Y} - \mathbf{X}\mathbf{b} = \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \\ &= [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = \mathbf{K}\mathbf{Y},\end{aligned} \tag{36}$$

unde $\mathbf{K}$ se bucură de proprietățile:

(37) $\mathbf{K}$ este simetrică și idempotentă (teorema X.2)

$$\mathbf{K}\mathbf{X} = 0. \tag{38}$$

DEFINIȚIA 8. Statistica

$$\mathrm{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) \tag{39}$$

poartă numele de suma pătratelor erorilor sau suma pătratelor reziduală

Avem

$$\mathrm{SSE} = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} \tag{40}$$

sau folosind $\mathbf{b}^*$ și $\mathscr{A}'\mathbf{Y}$

$$\begin{aligned}\mathrm{SSE} &= \mathbf{Y}'\mathbf{Y} - (\bar{Y} - \mathbf{b}^*\bar{\mathbf{X}}\ \ \mathbf{b}^{*\prime})\begin{pmatrix} n\bar{Y} \\ \mathbf{X}_1'\mathbf{Y}\end{pmatrix} = \\ &= \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2 - \mathbf{b}^{*\prime}(\mathbf{X}_1'\mathbf{Y} - n\bar{Y}\bar{\mathbf{X}})\end{aligned} \tag{40'}$$

Scanned with OKEN Scanner

TEOREMA 3. *Statistica*

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n - r(\mathbf{X})} = \frac{\text{SSE}}{n - r} \tag{41}$$

unde am folosit r pentru $r(\mathbf{X})$, *rangul lui* $\mathbf{X}$, *este un estimator nedeplasat pentru* σ^2.

Pentru demonstrație a se vedea teorema X.2

Observație. În cazul modelului de rang maxim $r = r(\mathbf{X}) = k + 1$.

DEFINIȚIA 9. Statistica

$$\text{SST} = \mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{n} Y_i^2 \tag{42}$$

poartă numele de suma totală a pătratelor.

DEFINIȚIA 10. Statistica

$$\text{SSR} = \text{SST} - \text{SSE} = \mathbf{b}'\mathbf{X}'\mathbf{Y} = \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} = n\bar{Y}^2 + \mathbf{b}^{*\prime}\mathscr{A}'\mathbf{Y} \tag{43}$$

poartă numele de suma pătratelor datorită regresiei sau suma redusă a pătratelor.

SSR reprezintă porțiunea din SST datorită ajustării regresiei.

Partiția SST poate fi reprezentată ca

$$\begin{array}{ll} \text{SSR} = \mathbf{b}'\mathbf{X}'\mathbf{Y} & = n\bar{Y}^2 + \mathbf{b}^{*\prime}\mathscr{A}'\mathbf{Y} \\ \text{SSE} = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} & = \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2 - \mathbf{b}^{*\prime}\mathscr{A}'\mathbf{Y} \\ \hline \text{SST} = \mathbf{Y}'\mathbf{Y} & = \mathbf{Y}'\mathbf{Y} \end{array} \tag{44}$$

În cele ce urmează vom considera modelul cu intercepție

$$Y_i = \beta_0 X_{i0} + \varepsilon_i = \beta_0 + \varepsilon_i.$$

Atunci $b_0 = \bar{Y}$ și $\text{SSR} = n\bar{Y}^2$ o vom numi corecția pentru medie și o vom scrie

$$\text{SSM} = n\bar{Y}^2. \tag{45}$$

Atunci din (44)

$$\text{SSR} = \text{SSM} + \mathbf{b}^{*\prime}\mathscr{A}'\mathbf{Y}$$

și are loc

DEFINIȚIA 11. Statistica

$$\text{SSR}_m = \text{SSR} - \text{SSM} = \mathbf{b}^{*\prime}\mathscr{A}'\mathbf{Y} = \mathbf{b}^{*\prime}\mathscr{A}'\mathscr{A}\mathbf{b}^*$$

poartă numele de suma pătratelor datorită regresiei corectată pentru medie.

Scanned with OKEN Scanner

Cu aceasta (44) devine

$$
\begin{aligned}
\text{SSM} &= n\bar{Y}^2 \\
\text{SSR}_m &= \mathbf{b}^{*\prime}\mathbf{A}'\mathbf{Y} = \mathbf{b}^{*\prime}\mathbf{A}'\mathbf{A}\mathbf{b}^* \\
\text{SSE} &= \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2 - \mathbf{b}^{*\prime}\mathbf{A}'\mathbf{Y} \\
\hline
\text{SST} &= \mathbf{Y}'\mathbf{Y}
\end{aligned}
\tag{44'}
$$

Similar cu SSR_m avem

DEFINIȚIA 12. Statistica

$$\text{SST}_m = \text{SST} - \text{SSM} = \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2 = y'y \tag{46}$$

poartă numele de suma totală a pătratelor corectată.

Avem

$$
\begin{aligned}
\text{SSR}_m &= \mathbf{b}^{*\prime}\mathbf{A}'\mathbf{Y} &&= \mathbf{b}^{*\prime}\mathbf{A}'\mathbf{Y} \\
\text{SSE} &= y'y - \mathbf{b}^{*\prime}\mathbf{A}'\mathbf{Y} &&= \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2 - \mathbf{b}^{*\prime}\mathbf{A}'\mathbf{Y} \\
\hline
\text{SST}_m &= y'y &&= \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2
\end{aligned}
\tag{47}
$$

PROPOZIȚIA 2. *Pătratul coeficientului de corelație dintre Y_i și valorile predicate $\hat{Y}_i$ este :*

— în modelul fără intercepție (media $\bar{Y}$ este ignorată)

$$R^2 = \frac{(\sum Y_i\hat{Y}_i)^2}{(\sum Y_i^2)(\sum \hat{Y}_i^2)} = \frac{(\mathbf{Y}'\hat{\mathbf{Y}})^2}{\mathbf{Y}'\mathbf{Y}(\hat{\mathbf{Y}}'\hat{\mathbf{Y}})} \tag{48}$$

sau

$$R^2 = \text{SSR}/\text{SST} \text{ și} \tag{49}$$

— în modelul cu intercepția

$$R^2 = \frac{[\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})]^2}{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{\hat{Y}})^2} \tag{50}$$

sau

$$R^2 = \text{SSR}_m/\text{SST}_m \tag{51}$$

Scanned with OKEN Scanner

2.4. Repartiția estimatorului lui β. Repartiții ale sumelor de pătrate

Teorema 4. *Dacă* $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, *atunci*

i) $$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$$

ii) $$\mathbf{b} \sim N[\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$$

ii') $$\mathbf{b}^* \sim N[\boldsymbol{\beta}^*, (\mathscr{A}'\mathscr{A})^{-1}\sigma^2)]$$

iii) $\mathbf{b}$ *și* $\hat{\sigma}^2$ *sînt independente*

iv) $$\mathrm{SSE}/\sigma^2 \sim \chi^2(n-r),\ r = \text{rang } (\mathbf{X}).$$

Demonstrație. Avem

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

și

$$\mathrm{SSE} = \mathbf{Y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\,\mathbf{Y} = \mathbf{Y}'\mathbf{K}\mathbf{Y}. \tag{52}$$

Din (38) avem

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{K} = \mathbf{0},$$

ceea ce demonstrează iii) (problema 11).

Din (52), $\mathrm{SSE} = \mathbf{Y}'\mathbf{K}\mathbf{Y}$ este o formă pătratică în $\mathbf{Y}$. Avem

$$\mathrm{SSE}/\sigma^2 = \mathbf{Y}'(1/\sigma^2)\,\mathbf{K}\mathbf{Y},$$

și după (37) $\mathbf{K}$ este idempotentă și $\mathscr{C}(\mathbf{Y}) = \sigma^2\mathbf{I}$. Prin urmare

$$(1/\sigma^2)\,\mathbf{K}\sigma^2\mathbf{I}$$

este idempotentă și (problema 11)

$$\mathrm{SSE}/\sigma^2 \sim \chi^2\{r[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'],\ \boldsymbol{\beta}'\mathbf{X}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\,\mathbf{X}\boldsymbol{\beta}/2\,\sigma^2\}$$

care datorită relațiilor (37) și (38) se reduce la

$$\mathrm{SSE}/\sigma^2 \sim \chi^2(n-r),\ r = r(\mathbf{X})$$

ceea ce demonstrează iv). Deci

$$(n-r)\,\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-r).$$

Teorema 5. *Dacă* $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$, *atunci*

i) $$\mathrm{SSR}/\sigma^2 \sim \chi^2(r, \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}/2\,\sigma^2)$$

ii) $$\mathrm{SSM}/\sigma^2 \sim \chi^2(1, (\mathbf{1}'\mathbf{X}\boldsymbol{\beta})^2/2\,n\sigma^2)$$

iii) $$\mathrm{SSR}_m/\sigma^2 \sim \chi^2(r-1, \boldsymbol{\beta}^{*\prime}\mathscr{A}'\mathscr{A}\boldsymbol{\beta}^*/2\,\sigma^2)$$

iv) $$\mathbf{Y}'\mathbf{Y}/\sigma^2 \sim \chi^2(n, \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}/2\,\sigma^2).$$

Scanned with OKEN Scanner

Demonstrație. Din (44) avem :

$$\text{SSR} = \mathbf{b}'\mathbf{X}'\mathbf{Y} = \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Matricea $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ este idempotentă (teorema X.2) și produsul ei cu matricea $\mathbf{K}$ definită în (36) este nul. Prin urmare, după problema 13, SSR este independentă de SSE și după problema 11

$$\text{SSR}/\sigma^2 \sim \chi^2\{r[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'], \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}/2\sigma^2\},$$

ceea ce demonstrează i).

Asemănător, din (44')

$$\text{SSM} = n\bar{Y}^2 = \mathbf{Y}'n^{-1}\mathbf{1}\mathbf{1}'\mathbf{Y}$$

unde $n^{-1}\mathbf{1}\mathbf{1}'$ este idempotentă și produsul său cu $\mathbf{K}$ este nul. Prin urmare SSM este repartizată independent de SSE și

$$\text{SSM}/\sigma^2 \sim \chi^2[r(n^{-1}\mathbf{1}\mathbf{1}'), \boldsymbol{\beta}'\mathbf{X}'n^{-1}\mathbf{1}\mathbf{1}'\mathbf{X}\boldsymbol{\beta}/2\sigma^2)$$

ceea ce demonstrează ii).

Din (47)

$$\text{SSR}_m = \mathbf{b}^{*\prime}\mathscr{A}'\mathbf{Y} = \mathbf{b}^{*\prime}\mathscr{A}'\mathscr{A}\mathbf{b}^*$$

și cum $\mathbf{b}^* \sim N[\boldsymbol{\beta}^*, (\mathscr{A}'\mathscr{A})^{-1}\sigma^2]$ (teorema 4, pct. ii'),

$$\text{SSR}_m/\sigma^2 \sim \chi^2[r(\mathscr{A}'\mathscr{A}), \boldsymbol{\beta}^{*\prime}\mathscr{A}'\mathscr{A}\boldsymbol{\beta}^*/2\sigma^2],$$

ceea ce demonstrează iii).

Mai mult, SSR_m poate fi exprimată ca $\mathbf{Y}'\mathbf{Q}\mathbf{Y}$ unde $\mathbf{Q}$ este nu numai idempotentă însă produsele ei cu $\mathbf{K}$ și $n^{-1}\mathbf{1}\mathbf{1}'$ sînt nule. Deci după problema 13, SSR_m este independentă atît de SSE cît și de SSM.

COROLARUL *3. Avem* :

$$(53) \qquad F(R) = \frac{\text{SSR}/r}{\text{SSE}/(n-r)} \sim F(r, n-r, \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}/2\sigma^2)$$

$$(54) \qquad F(M) = \frac{\text{SSM}/1}{\text{SSE}/(n-r)} \sim F(1, n-r, (\mathbf{1}'\mathbf{X}\boldsymbol{\beta})^2/2n\sigma^2)$$

și

$$(55) \qquad F(R_m) = \frac{\text{SSR}_m/(r-1)}{\text{SSE}/(n-r)} \sim F(r-1, n-r, \boldsymbol{\beta}^{*\prime}\mathscr{A}'\mathscr{A}\boldsymbol{\beta}^*/2\sigma^2).$$

Scanned with OKEN Scanner

Corolarul 3 permite întocmirea următoarelor tabele

Tabelul 1

Analiza dispersiei pentru regresia ajustată

Sursa variației	Suma pătratelor (SS)	Gr. de libertate	Media sumei pătratelor ($\overline{SS}$)	Statistica F
Regresia	$SSR = \mathbf{b}'\mathbf{X}'\mathbf{Y}$	r	$\overline{SSR} = SSR/r$	$F(R) = \dfrac{\overline{SSR}}{\overline{SSE}}$
Eroarea reziduală	$SSE = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$	$n - r$	$\overline{SSE} = SSE/(n-r)$	
Total	$SST = \mathbf{Y}'\mathbf{Y}$	n		

Tabelul 2

Analiza dispersiei, cu un termen pentru medie

Sursa variației	Suma pătratelor (SS)	Gr. de libertate	$\overline{SS}$	Statistica F
Media	$SSM = n\bar{Y}^2$	1	$\overline{SSM} = \dfrac{SSM}{1}$	$F(M) = \dfrac{\overline{SSM}}{\overline{SSE}}$
Regresia corectată ptr. medie	$SSR_m = \mathbf{b}^{*\prime}\mathbf{A}'\mathbf{Y}$	$r^{*)} - 1$	$\overline{SSR}_m = \dfrac{SSR_m}{r-1}$	$F(R_m) = \dfrac{\overline{SSR}_m}{\overline{SSE}}$
Eroarea reziduală	$SSE = \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2 - \mathbf{b}^{*\prime}\mathbf{A}'\mathbf{Y}$	$n - r$	$\overline{SSE} = \dfrac{SSE}{n-r}$	
Total	$SST = \mathbf{Y}'\mathbf{Y}$	n		

*) $r = r(\mathbf{X}) = k + 1$

Tabelul 3

Analiza dispersiei (corectată pentru medie)

Sursa variației	Suma pătratelor	Gr. de libert.	$\overline{SS}$	Statistica F
Regresia corectată ptr. medie	$SSR_m = \mathbf{b}^{*\prime}\mathbf{A}'\mathbf{Y}$	$r - 1$	$\overline{SSR}_m = \dfrac{SSR_m}{r-1}$	$F(R_m) = \dfrac{\overline{SSR}_m}{\overline{SSE}}$
Eroarea reziduală	$SSE = \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2 - \mathbf{b}^{*\prime}\mathbf{A}'\mathbf{Y}$	$n - r$	$\overline{SSE} = \dfrac{SSE}{n-r}$	
Total corectat ptr. medie	$SST_m = \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2$	$n - 1$		

Scanned with OKEN Scanner

2.5. Eroarea pură

Presupunem că sînt p valori distincte ale lui X

$$X_1 \; X_2, \ldots, X_p$$

unde X_i apare de n_i ori, adică cu n_i valori Y_{ij} pentru $j = 1, \ldots, n_i$ și pentru $i = 1, 2, \ldots, p$. Pentru toți i, $n_i \geqslant 1$ și vom scrie

$$n_\cdot = \sum_{i=1}^{p} n_i = n.$$

Atunci

$$\text{SSE} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} Y_{ij}^2 - \mathbf{b}'\mathbf{X}'\mathbf{Y} =$$

(56)

$$= \sum_{i=1}^{p} \sum_{j=1}^{n_i} Y_{ij}^2 - nY_{\cdot\cdot}^2 - b^* \left(\sum_{i=1}^{p} \sum_{j=1}^{n_i} X_{ij} Y_{ij} - n\bar{X}_{\cdot\cdot}\bar{Y}_{\cdot\cdot} \right)$$

cu $n - 2$ grade de libertate, poate fi partiționată în

(57) $$\text{SS}_{ep} = \sum_{i=1}^{p} \left[\sum_{j=1}^{n_i} Y_{ij}^2 - n_i(\bar{Y}_{i\cdot})^2 \right]$$ cu $n - p$ grade de libertate

și

$$\text{SS}_{la} = \text{SSE} - \text{SS}_{ep} \text{ cu } p - 2 \text{ grade de libertate.}$$

DEFINIȚIA 13. Statistica

(59) $$\text{SS}_{ep}/(n - p)$$

poartă numele de media sumei pătratelor datorită erorii pure.

Statistica

(60) $$\text{SS}_{la}/(p - 2)$$

poartă numele de media sumei pătratelor datorită lipsei de ajustare.

Un test pentru lipsa de ajustare se obține comparînd

(61) $$F(la) = \frac{\text{SS}_{la}/(p - 2)}{\text{SS}_{ep}/(n - p)} \sim F(p - 2, n - p)$$

cu $F_{1-\alpha}(p - 2, n - p)$. Semnificația indică faptul că modelul nu este corespunzător. Lipsa de semnificație indică că nu avem motive să ne îndoim că modelul este corespunzător, în care caz, $\text{SSE}/(n - 2)$ este un estimator ponderat pentru σ^2.

Scanned with OKEN Scanner

2.6. Verificarea ipotezelor

În tabelul 1, statistica $F(R)$ este repartizată F necentrat cu parametrul de necentricitate $\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}/2\sigma^2$ (53). Acest parametru este zero dacă este adevărata ipoteză nulă

$$H : \boldsymbol{\beta} = \mathbf{0}$$

cînd $F(R)$ are o repartiție F centrată ($F(r, n - r)$).

Respingem ipoteza H cînd

$$F(R) > F_{1-\alpha}(r, n - r).$$

În tabelul 2, statistica $F(M)$ este repartizată F necentrat cu parametrul de necentricitate $(\mathbf{1}'\mathbf{X}\boldsymbol{\beta})^2/2n\sigma^2$. Numărătorul acestei expresii se scrie

$$\mathbf{1}'\mathbf{X}\boldsymbol{\beta} = \mathbf{1}'M(\mathbf{Y}) = M(\mathbf{1}'\mathbf{Y}) = M(n\bar{Y}) = nM(\bar{Y}).$$

Deci parametrul de necentricitate în (54) este $n[M(\bar{Y})]^2/2\sigma^2$. Acest parametru este zero dacă este adevărată ipoteza

$$H : M(\bar{Y}) = 0;$$

cînd $F(M)$ are o repartiție F centrată ($F(1, n - r)$).

$F(M)$ poate fi folosită pentru verificarea ipotezei conform căreia valoarea medie a mediei Y-lor observați este zero.

Echivalent $\sqrt{F(M)} \sim t(n - r)$, deoarece

$$F(M) = n\bar{Y}^2/\hat{\sigma}^2 = [\bar{Y}/(\hat{\sigma}/\sqrt{n})]^2$$

este pătratul unei variabile t.

Un alt test care folosește $F(M)$ este bazat pe modelul $M(Y_i) = \beta_0$.

Suma pătratelor redusă pentru ajustarea acestui model este SSM și atunci parametrul de necentricitate din (54) este $n\beta_0^2/2\sigma^2$.

Deci $F(M)$ poate fi folosit pentru a verifica dacă modelul $M(Y_i) = \beta_0$ ia în considerație variația în variabila Y.

Testul bazat pe $F(R)$ verifică ipoteza conform căreia toți β_i, inclusiv β_0 sînt simultan zero. Totuși pentru verificarea ipotezei nule

$$H : \boldsymbol{\beta}^* = \mathbf{0}$$

se folosește statistica $F(R_m)$ din tabelele 2 și 3. Aceasta rezultă din faptul că parametrul de necentricitate din repartiția F necentrată a statisticii $F(R_m)$ este zero cînd este adevărată ipoteza $\boldsymbol{\beta}^* = \mathbf{0}$, caz în care $R(R_m) \sim$ $\sim F(r - 1, n - r)$. Astfel $F(R_m)$ stă la baza unui test pentru ipoteza $\boldsymbol{\beta}^* = \mathbf{0}$. Dacă $F(R_m)$ este semnificativ ipoteza este respinsă. Aceasta nu înseamnă că toate elementele lui β^* sînt nenule, ci că cel puțin unul dintre ele poate fi nenul. Dacă a fost găsit că $F(M)$ este semnificativ, atunci $F(R_m)$ fiind semnificativ indică că modelul conținînd X-ișii explică mai mult variația în variabilă Y decît modelul $M(Y) = \beta_0$.

Scanned with OKEN Scanner

Observație. Testele ce folosesc $F(M)$ și $F(R_m)$ sînt bazate pe SSM și SSR_m, statistici ce sînt independente. Prin urmare semnificația sau nesemnificația statisticii $F(M)$ și/sau $F(R_m)$ este independentă de numărătorul celcilalte, cu toate că statisticile F nu sînt independente deoarece au același numitor.

2.7. Intervale de încredere

PROPOZIȚIA 3. *Avem*

i) $$\frac{b_i - \beta_i}{\sqrt{a^{ii}\sigma^2}} \sim N(0, 1), \quad i = 0, 1, \ldots, k$$

unde, a^{ii} *este element diagonal de ordin* i *al matricei* $(\mathbf{X}'\mathbf{X})^{-1}$

$a^{00} = 1/n + \overline{\mathbf{X}}'(\mathscr{A}'\mathscr{A})^{-1}\overline{\mathbf{X}}$ *(din* (26) *și* (27) *și*

a^{ii} = *elementul diagonal de ordin* i *al matricei* $(\mathscr{A}'\mathscr{A})^{-1}$

ii) $$\frac{b_i - \beta_i}{\sqrt{a^{ii}\hat{\sigma}^2}} \sim t(n - r)$$

unde $\hat{\sigma}^2$ *este dat de* (41)

iii) $$\frac{\mathbf{c}'\mathbf{b} - \mathbf{c}'\boldsymbol{\beta}}{\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}\hat{\sigma}^2}} \sim t(n - r).$$

Demonstrația rezultă din teorema 4.

COROLARUL 4. *Intervalul*

(62) $$(b_i - \hat{\sigma}t_{1-(\alpha/2)}(n - r)\sqrt{a^{ii}}, \quad b_i + \hat{\sigma}t_{1-(\alpha/2)}(n - r)\sqrt{a^{ii}})$$

este un interval de încredere pentru β_i *avînd coeficientul de încredere* $1 - \alpha$.

Demonstrația rezultă din faptul că statistica (ii) satisface condițiile a), b) și c) din § 2 capitolul XII pentru existența intervalelor de încredere.

COROLARUL 5. *Cînd numărul gradelor de libertate este mare (de exemplu* $n - r > 100$)

(63) $$(b_i - \hat{\sigma}z_{1-(\alpha/2)}\sqrt{a^{ii}}, \quad b_i + \hat{\sigma}z_{1-(\alpha/2)}\sqrt{a^{ii}})$$

este un interval de încredere pentru β_i *avînd coeficientul de încredere* $1 - \alpha$.

Demonstrația rezultă din faptul că repartiția statisticii ii) este aproximativ $N(0, 1)$ (§ 3 cap. I).

COROLARUL 6. *Intervalul*

(64) $$(\mathbf{c}'\mathbf{b} - \hat{\sigma}t_{1-(\alpha/2)}(n - r)\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}, \quad \mathbf{c}'\mathbf{b} +$$

$$+ \hat{\sigma}t_{1-(\alpha/2)}(n - r)\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}})$$

este un interval de încredere pentru $\mathbf{c}'\boldsymbol{\beta}$ *avînd coeficientul de încredere* $1 - \alpha$.

Scanned with OKEN Scanner

Demonstrația rezultă din faptul că statistica iii) satisface condițiile a), b) și c) din § 2 capitolul XII pentru existența intervalelor de încredere.

Din (30) și (64) obținem

COROLARUL 7. *Intervalul*

$$(65)\qquad (\mathbf{X}_0'\mathbf{b} - \hat{\sigma}t_{1-(\alpha/2)}(n-r)\sqrt{\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0},\ \mathbf{X}_0'\mathbf{b} + \hat{\sigma}t_{1-(\alpha/2)}(n-r)\sqrt{\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0})$$

este un interval de încredere pentru $\mathbf{X}_0'\boldsymbol{\beta}$ *având coeficientul de încredere* $1-\alpha$.

Observații. i) În cazul regresiei cu o singură variabilă $X(k=1$, și $r=2)$, $\mathbf{X}_0' = (1\ X_0)$ și (65) devine

$$\left((1\ X_0)\begin{pmatrix}\bar{Y}-b\bar{X}\\ b\end{pmatrix} - \hat{\sigma}t_{1-(\alpha/2)}(n-2)\sqrt{(1\ X_0)\begin{pmatrix} n & n\bar{X}\\ n\bar{X} & \sum_1^n X_i^2\end{pmatrix}^{-1}\begin{pmatrix}1\\ X_0\end{pmatrix}},\right.$$

$$\left.(1\ X_0)\begin{pmatrix}\bar{Y}-b\bar{X}\\ b\end{pmatrix} + \hat{\sigma}t_{1-(\alpha/2)}(n-2)\sqrt{(1\ X_0)\begin{pmatrix} n & n\bar{X}\\ n\bar{X} & \sum_1^n X_i^2\end{pmatrix}^{-1}\begin{pmatrix}1\\ X_0\end{pmatrix}}\right)$$

sau

$$\left(\bar{Y} - b(X_0-\bar{X}) - \hat{\sigma}t_{1-(\alpha/2)}(n-2)\sqrt{\frac{1}{n}+\frac{(\bar{X}-X_0)^2}{\sum_{i=1}^n X_i^2 - n\bar{X}^2}},\right.$$

$$(66)$$

$$\left.\bar{Y} - b(X_0-\bar{X}) + \hat{\sigma}t_{1-(\alpha/2)}(n-2)\sqrt{\frac{1}{n}+\frac{(\bar{X}-X_0)^2}{\sum_{i=1}^n X_i^2 - n\bar{X}^2}}\right),$$

ceea ce constituie un interval de încredere pentru $M(Y)$. Reprezentînd limitele acestui interval pentru o serie de valori ale lui X_0, obținem o bandă de încredere pentru dreapta de regresie $Y = \beta_0 + \beta_1 X$

ii) Ținînd seama de (34) în (65), adică înlocuind

$$\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0 \text{ prin } \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0 + 1$$

obținem în cazul regresiei simple următorul interval de toleranță

$$\left(\bar{Y} + b(X_0-\bar{X}) - \hat{\sigma}t_{1-(\alpha/2)}(n-r)\sqrt{1+\frac{1}{n}+\frac{(\bar{X}-X_0)}{\sum_{i=1}^n X_i^2 - n\bar{X}^2}},\right.$$

$$(67)$$

$$\left.\bar{Y} + b(X_0-\bar{X}) + \hat{\sigma}t_{1-(\alpha/2)}(n-r)\sqrt{1+\frac{1}{n}+\frac{\bar{X}-X_0}{\sum_{i=1}^n X_i^2 - n\bar{X}^2}}\right)$$

Scanned with OKEN Scanner

§ 3. Ipoteza liniară generală

3.1. Verificarea ipotezei liniare generale

Prezintă intens verificarea următoarelor 4 ipoteze :

i) $H : \boldsymbol{\beta} = \mathbf{0}$, ipoteza conform căreia toate elementele vectorului $\boldsymbol{\beta}$ sînt zero

ii) $H : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, ipoteza conform căreia $\beta_i = \beta_{i0}$, $i = 0, 1, \ldots, k$ o valoare specificată

iii) $H : \boldsymbol{\lambda}'\boldsymbol{\beta} = \mathbf{m}$, ipoteza conform căreia o anumită combinație liniară de elementele lui $\boldsymbol{\beta}$ este egală cu o anumită constantă dată.

iv) $H : \boldsymbol{\beta}_q = \mathbf{0}$, ipoteza conform căreia $q (q < k)$ elemente ale lui $\boldsymbol{\beta}$ sînt zero.

Toate aceste sînt cazuri speciale ale ipotezei liniare generale

$$H : \mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$$

unde : $\boldsymbol{\beta}$ este vectorul parametrilor modelului (de ordin $k + 1$)
$\mathbf{c}'$ este orice matrice $s \times (k + 1)$ și
$\mathbf{m}$ este un vector de constante specificate de ordin s.

Asupra matricii $\mathbf{C}'$ avem o singură limitare : să aibe rangul maxim egal cu numărul de linii, adică de $r(\mathbf{C}') = s$. Aceasta arată că funcțiile liniare de $\boldsymbol{\beta}$ care formează ipoteza trebuie să fie liniar independente. De asemenea se cere ca $\mathbf{m}$ să fie astfel încît ecuațiile $\mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$ să fie consistente. Aceasta se atinge cînd $\mathbf{C}'$ este de rang maxim s, caz în care ecuațiile $\mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$ sînt consistente pentru orice vector $\mathbf{m}$.

TEOREMA 6. *Avem*

$$\mathbf{C}'\mathbf{b} - \mathbf{m} \sim N[\mathbf{C}'\boldsymbol{\beta} - \mathbf{m}, \mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}\sigma^2]$$

Demonstrația rezultă din teorema 4.

TEOREMA 7. *Dacă*

$$Q = (\mathbf{C}'\mathbf{b} - \mathbf{m})' (\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C})^{-1}(\mathbf{C}'\mathbf{b} - \mathbf{m}) \tag{68}$$

atunci

$$Q/\sigma^2 \sim \chi^2[s, (\mathbf{C}'\boldsymbol{\beta} - \mathbf{m}) (\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C})^{-1}(\mathbf{C}'\boldsymbol{\beta} - \mathbf{m})/2\sigma^2] \tag{69}$$

și Q *și* SSE *sînt independente.*

Demonstrație. Deoarece Q este o formă patratică în $\mathbf{C}'\mathbf{b} - \mathbf{m}$ (repartizată $N[\mathbf{C}'\boldsymbol{\beta} - \mathbf{m}, \mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}\sigma^2]$ după teorema anterioară) cu matricea $(\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C})^{-1}$, aplicarea problemei 11 demonstrează (69).

Folosim problema 13 pentru a arăta că Q și SSE sînt independente. Mai întîi vom arăta că Q și SSE se exprimă ca forme pătratice de aceeași variabilă normală. Deoarece $\mathbf{C}'$ are rangul egal cu numărul de linii și $(\mathbf{X}'\mathbf{X})$

Scanned with OKEN Scanner

este simetrică, există inversa matricei $\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}$ folosite în (69). Înlocuind $\mathbf{b}$ din expresia lui Q prin $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, obținem

$$Q = [\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - \mathbf{m}]' [\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1} [\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - \mathbf{m}].$$

Cum $\mathbf{C}'$ are rangul egal cu numărul de linii, după problema 9, există $(\mathbf{C}'\mathbf{C})^{-1}$. Prin urmare

$$\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - \mathbf{m} = \mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' [\mathbf{Y} - \mathbf{X}\mathbf{C}(\mathbf{C}'\mathbf{C})^{-1}\mathbf{m}]$$

și astfel

$$Q = [\mathbf{Y} - \mathbf{X}\mathbf{C}(\mathbf{C}'\mathbf{C})^{-1}\mathbf{m}]' \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1} \times$$

$$\times \mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' [\mathbf{Y} - \mathbf{X}\mathbf{C}(\mathbf{C}'\mathbf{C})^{-1}\mathbf{m}].$$

Acum considerăm suma pătratelor erorilor

$$\mathrm{SSE} = \mathbf{Y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \mathbf{Y} = \mathbf{Y}'\mathbf{K}\mathbf{Y}.$$

Deoarece $\mathbf{X}'\mathbf{K}$ și $\mathbf{K}\mathbf{X}$ sînt amîndouă nule, SSE poate fi scris ca

$$\mathrm{SSE} = [\mathbf{Y} - \mathbf{X}\mathbf{C}(\mathbf{C}'\mathbf{C})^{-1}\mathbf{m}]' \mathbf{K}[\mathbf{Y} - \mathbf{X}\mathbf{C}(\mathbf{C}'\mathbf{C})^{-1}\mathbf{m}].$$

Atît Q cît și SSE au fost exprimate ca forme pătratice în vectorul $\mathbf{Y} - \mathbf{X}\mathbf{C}(\mathbf{C}'\mathbf{C})^{-1}\mathbf{m}$, care este normal repartizat, matricea în fiecare formă pătratică fiind idempotentă .Produsul celor două matrici este nul.

$$\mathbf{K}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{0}$$

Prin urmare, după problema 13, Q și SSE sînt repartizate independent. Teorema este demonstrată.

COROLARUL 8. *Statistica*

$$F(H) = \frac{Q/s}{\mathrm{SSE}/[n - r(\mathbf{X})]} = Q/s\hat{\sigma}^2 \tag{70}$$

este repartizată

$$F\{s, n - r(\mathbf{X}), (\mathbf{C}'\boldsymbol{\beta} - \mathbf{m})' [\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}(\mathbf{C}'\boldsymbol{\beta} - \mathbf{m})/2\sigma^2\}$$

și în ipoteza nulă $H: \mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$

$$F(H) \sim F(s, n - r(\mathbf{X})).$$

Prin urmare statistica $F(H)$ stă la baza unui test pentru verificarea ipotezei H. În această situație

$$F(H) = \frac{Q}{s\hat{\sigma}^2} = \frac{(\mathbf{C}'\mathbf{b} - \mathbf{m})' [\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1} (\mathbf{C}'\mathbf{b} - \mathbf{m})}{s\hat{\sigma}^2} \tag{71}$$

și $F(H) \sim F(s, n - r)$, s fiind numărul liniilor matricii $\mathbf{C}'$, care are rangul maxim egal cu numărul de linii.

Scanned with OKEN Scanner

Acest test este general: el se aplică pentru orice ipoteză liniară $\mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$ singura condiție fiind: $\mathbf{C}'$ să aibe rangul maxim egal cu numărul de linii. $\hat{\sigma}^2$ este universal în orice aplicație a statisticii $F(H)$. Astfel considerînd diferite ipoteze singurul termen din $F(H)$ care se schimbă este Q/s.

Un test de semnificație respinge ipoteza $H: \mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$ la pragul de semnificație α, ori de cîte ori $F(H) > F_{1-\alpha}(s, n-r)$.

3.2. Estimația în ipoteza nulă

Considerînd ipoteza $H: \mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$ se ridică problema: Care este estimatorul parametrului $\boldsymbol{\beta}$ în ipoteza nulă? Această problemă este importantă în cazul nerespingerii ipotezei cu ajutorul testului bazat pe statistica F. Estimatorul $\tilde{\mathbf{b}}$ căutat se obține folosind C.M.M.P. cu constrîngeri: Astfel $\tilde{\mathbf{b}}$ se obține minimizînd $(\mathbf{Y} - \mathbf{X}\tilde{\mathbf{b}})'(\mathbf{Y} - \mathbf{X}\tilde{\mathbf{b}})$ cu condiția $\mathbf{C}'\tilde{\mathbf{b}} = \mathbf{m}$. Notînd cu $2\boldsymbol{\theta}'$, multiplicatorul lui Lagrange minimizăm

$$(\mathbf{Y} - \mathbf{X}\tilde{\mathbf{b}})'(\mathbf{Y} - \mathbf{X}\tilde{\mathbf{b}}) + 2\boldsymbol{\theta}'(\mathbf{C}'\tilde{\mathbf{b}} - \mathbf{m}) \tag{72}$$

în raport cu elementele lui $\tilde{\mathbf{b}}$ și $\boldsymbol{\theta}$. Derivînd (72) în raport cu aceste elemente obținem ecuațiile

$$\begin{cases} \mathbf{X}'\mathbf{X}\tilde{\mathbf{b}} + \mathbf{C}\boldsymbol{\theta} = \mathbf{X}'\mathbf{Y} \\ \mathbf{C}'\tilde{\mathbf{b}} = \mathbf{m}. \end{cases} \tag{73}$$

Din prima ecuație găsim

$$\tilde{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y} - \mathbf{C}\boldsymbol{\theta}) = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}\boldsymbol{\theta}$$

și din a doua

$$\mathbf{C}'\tilde{\mathbf{b}} = \mathbf{C}'\mathbf{b} - \mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}\boldsymbol{\theta} = \mathbf{m}.$$

Deci

$$\boldsymbol{\theta} = [\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}(\mathbf{C}'\tilde{\mathbf{b}} - \mathbf{m})$$

și astfel

$$\tilde{\mathbf{b}} = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}[\mathbf{C}'(\mathbf{X}'\mathbf{X})\mathbf{C}]^{-1}(\mathbf{C}'\mathbf{b} - \mathbf{m}). \tag{74}$$

Am demonstrat

PROPOZIȚIA 4. *In ipoteza* $H: \mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$, *estimatorul de* C.M.M.P. *pentru* $\boldsymbol{\beta}$ *este dat de* (74).

PROPOZIȚIA 5. *In ipoteza* $H: \mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$, *suma pătratelor reziduriloi este*

$$\mathrm{SSE} + Q, \tag{75}$$

unde Q *este numărătorul statisticii* $F(H)$ *(din* 71)).

Scanned with OKEN Scanner

Demonstrație. Suma pătratelor rezidurilor este

$$(\mathbf{Y}-\mathbf{X}\tilde{\mathbf{b}})'(\mathbf{Y}-\mathbf{X}\tilde{\mathbf{b}}) = [\mathbf{Y}-\mathbf{X}\mathbf{b}+\mathbf{X}(\mathbf{b}-\tilde{\mathbf{b}})]'[\mathbf{Y}-\mathbf{X}\mathbf{b}+\mathbf{X}(\mathbf{b}-\tilde{\mathbf{b}})] =$$

(76)

$$= (\mathbf{Y}-\mathbf{X}\mathbf{b})'(\mathbf{Y}-\mathbf{X}\mathbf{b}) + (\tilde{\mathbf{b}}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\tilde{\mathbf{b}}-\mathbf{b}),$$

ceilalți termeni anulîndu-se deoarece $\mathbf{X}'(\mathbf{Y}-\mathbf{X}\mathbf{b}) = \mathbf{0}$. Din (74) obținem:

$$\mathbf{b}-\tilde{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}(\mathbf{C}'\mathbf{b}-\mathbf{m}).$$

Cu aceasta (76), devine

$$\begin{aligned}(\mathbf{Y}-\mathbf{X}\tilde{\mathbf{b}})'(\mathbf{Y}-\mathbf{X}\tilde{\mathbf{b}}) &= \mathrm{SSE} + \\ &+ (\mathbf{C}'\mathbf{b}-\mathbf{m})'[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\times \\ &\times \mathbf{C}[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}(\mathbf{C}'\mathbf{b}-\mathbf{m}) = \\ &= \mathrm{SSE} + (\mathbf{C}'\mathbf{b}-\mathbf{m})'[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}(\mathbf{C}'\mathbf{b}-\mathbf{m}),\end{aligned} \tag{77}$$

ceea ce demonstrează propoziția.

3.3. Aplicații

Folosim expresiile pentru $F(H)$ (71) și $\tilde{\mathbf{b}}$ (74) pentru cele 4 ipoteze din 3.1.

i) $H: \boldsymbol{\beta} = \mathbf{0}$. Scriem $\boldsymbol{\beta} = \mathbf{0}$ ca $\mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$, luînd

$$\mathbf{C}' = \mathbf{I},\ s = k+1 \text{ și } \mathbf{m} = 0.$$

Astfel $[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1} = \mathbf{X}'\mathbf{X}$
și

$$F(H) = \frac{\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}}{(k+1)\hat{\sigma}^2} = \frac{\mathrm{SSR}}{r}\cdot\frac{n-r}{\mathrm{SSE}}. \tag{78}$$

În ipoteza nulă $F(H) \sim F(r, n-r)$, unde $r = k+1$.
Valoarea corespunzătoare $\tilde{\mathbf{b}}$ este

$$\tilde{\mathbf{b}} = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}[(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\mathbf{b} = \mathbf{0}.$$

Pentru verificarea ipotezei $\boldsymbol{\beta}^* = \mathbf{0}$, statistica F este

$$\mathrm{SSR}_m/k\hat{\sigma}^2 \tag{79}$$

Scanned with OKEN Scanner

Scanned with OKEN Scanner

și $\tilde{\mathbf{b}}$ are expresia

$$\tilde{\mathbf{b}}' = (\bar{Y}\ \ \mathbf{0}'). \tag{80}$$

ii) $H : \boldsymbol{\beta} = \boldsymbol{\beta}_0$. Scriem $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ ca $\mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$ luînd

$$\mathbf{C}' = \mathbf{I},\ s = k + 1,\ \mathbf{m} = \boldsymbol{\beta}_0.$$

Astfel $[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1} = \mathbf{X}'\mathbf{X}$

și

$$F(H) = \frac{(\mathbf{b} - \boldsymbol{\beta}_0)'\mathbf{X}'\mathbf{X}(\mathbf{b} - \boldsymbol{\beta}_0)}{(k+1)\,\hat{\sigma}^2}. \tag{81}$$

Putem să scriem

$$(\mathbf{b} - \boldsymbol{\beta}_0)'\mathbf{X}'\mathbf{X}(\mathbf{b} - \boldsymbol{\beta}_0) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0) =$$

$$= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0).$$

În ipoteza nulă $F(H) \sim F(r, n - r)$, unde $r = k + 1$.

Valoarea corespunzătoare $\tilde{\mathbf{b}}$ (estimatorul lui $\boldsymbol{\beta}$) este $\tilde{\mathbf{b}} = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}[(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\mathbf{b} - \boldsymbol{\beta}_0) = \boldsymbol{\beta}_0$.
Pentru verificarea ipotezei $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0^*$ statistica F este

$$(\mathbf{b}^* - \boldsymbol{\beta}_0^*)'\mathscr{A}'\mathscr{A}(\mathbf{b}^* - \boldsymbol{\beta}_0^*)/k\hat{\sigma}^2 \tag{82}$$

și

$$\tilde{\mathbf{b}}' = (\bar{Y} - \bar{\mathbf{X}}'\boldsymbol{\beta}_0^*\ \ \boldsymbol{\beta}_0^{*\prime}). \tag{83}$$

iii) $H : \boldsymbol{\lambda}'\boldsymbol{\beta} = m$. Scriem $\boldsymbol{\lambda}'\boldsymbol{\beta} = m$ ca $\mathbf{C}\boldsymbol{\beta}' = \mathbf{m}$ luînd

$$\mathbf{C}' = \boldsymbol{\lambda}',\ s = 1,\ \mathbf{m} = m.$$

Astfel

$$F(H) = \frac{(\boldsymbol{\lambda}'\mathbf{b} - m)'[\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\lambda}]^{-1}(\boldsymbol{\lambda}'\mathbf{b} - m)}{\hat{\sigma}^2} \tag{84}$$

sau

$$F(H) = \frac{(\boldsymbol{\lambda}'\mathbf{b} - m)^2}{\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\lambda}\hat{\sigma}^2}. \tag{84'}$$

În ipoteza nulă $F(H) \sim F(1, n - r)$. Deci

$$\sqrt{F(H)} = \frac{\boldsymbol{\lambda}'\mathbf{b} - m}{\hat{\sigma}\sqrt{\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\lambda}}} \sim t(n - r) \tag{85}$$

deoarece $\lambda'\mathbf{b}$ este repartizat normal cu dispersia $\lambda'(\mathbf{X}'\mathbf{X})^{-1}\lambda\sigma^2$. Pentru această ipoteză

$$\tilde{\mathbf{b}} = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\lambda[\lambda'(\mathbf{X}'\mathbf{X})^{-1}\lambda]^{-1}(\lambda'\mathbf{b} - m) =$$

$$= \mathbf{b} - \left\{\frac{\lambda'\mathbf{b} - m}{\lambda'(\mathbf{X}'\mathbf{X})^{-1}\lambda}\right\}(\mathbf{X}'\mathbf{X})^{-1}\lambda.$$

Pentru verificarea ipotezei $\lambda'\boldsymbol{\beta}^* = \mathbf{m}$ statistica F este

$$(\lambda'\mathbf{b}^* - \mathbf{m})^2/\lambda'(\mathscr{A}'\mathscr{A})^{-1}\lambda\hat{\sigma}^2 \tag{86}$$

și

$$\tilde{\mathbf{b}} = \mathbf{b} + \left(\frac{\lambda'\mathbf{b}^* - \mathbf{m}}{\lambda'(\mathscr{A}'\mathscr{A})^{-1}\lambda}\right)\begin{pmatrix}\overline{\mathbf{X}}'\\ -\mathbf{I}\end{pmatrix}(\mathscr{A}'\mathscr{A})^{-1}\lambda. \tag{87}$$

iv) $H: \boldsymbol{\beta}_q = \mathbf{0}$, adică $b_i = 0$ pentru $i=0, 1, \ldots, q-1$, pentru $q<k$. În acest caz

$$\mathbf{C}' = (\mathbf{I}_q\ \mathbf{0}) \text{ și } \mathbf{m} = \mathbf{0} \text{ astfel încît } s = q.$$

Fie

$$\boldsymbol{\beta}_q' = (\beta_0\beta_1 \ldots \beta_{q-1})$$

și considerăm partițiile:

$$\boldsymbol{\beta} = \begin{pmatrix}\boldsymbol{\beta}_q\\ \boldsymbol{\beta}_p\end{pmatrix},\quad \mathbf{b} = \begin{pmatrix}\mathbf{b}_q\\ \mathbf{b}_p\end{pmatrix} \text{ și } (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix}\mathbf{U}_{qq} & \mathbf{U}_{qp}\\ \mathbf{U}_{pq} & \mathbf{U}_{pp}\end{pmatrix}$$

unde $p + q =$ ordinul lui $\boldsymbol{\beta} = k + 1$. Atunci

$$\mathbf{C}'\mathbf{b} = \mathbf{b}_q$$

$$[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1} = \mathbf{U}_{qq}^{-1}$$

și (71) devine

$$F(H) = \mathbf{b}_q'\mathbf{U}_{qq}^{-1}\mathbf{b}_q/q\hat{\sigma}^2. \tag{88}$$

Cu toate că rezultatul obținut se referă la primele q componente ale lui $\boldsymbol{\beta}$, el se aplică la orice submulțime de q componente ale lui $\boldsymbol{\beta}$. În particular pentru un singur β, se obține testul F cu un grad de libertate, echivalent cu testul t.

Sub această ipoteză

$$\tilde{\mathbf{b}} = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\begin{pmatrix}\mathbf{I}_q\\ \mathbf{0}\end{pmatrix}\mathbf{U}_{qq}^{-1}(\mathbf{b}_q - \mathbf{0}) = \mathbf{b} - \begin{pmatrix}\mathbf{U}_{qq}\\ \mathbf{U}_{pq}\end{pmatrix}\mathbf{U}_{qq}^{-1}\mathbf{b}_q =$$

$$= \begin{pmatrix}\mathbf{b}_q\\ \mathbf{b}_p\end{pmatrix} - \begin{pmatrix}\mathbf{b}_q\\ \mathbf{U}_{pq}\mathbf{U}_{qq}^{-1}\mathbf{b}_q\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ \mathbf{b}_p - \mathbf{U}_{pq}\mathbf{U}_{qq}^{-1}\mathbf{b}_q\end{pmatrix}.$$

Scanned with OKEN Scanner

Pentru verificarea ipotezei $\beta_q^* = 0$ statistica F este

(89) $$\mathbf{b}_q^{*\prime}\mathcal{U}_{qq}^{-1}\mathbf{b}_q^*/q\hat{\sigma}^2$$

şi

(90) $$\tilde{\mathbf{b}} = \begin{pmatrix} \bar{Y} - \bar{\mathbf{X}}_p'\tilde{\mathbf{b}}_p^* \\ 0 \end{pmatrix}$$

(91) $$\tilde{\mathbf{b}}_q^* = \mathbf{b}_p^* - \mathcal{U}_{pq}\mathcal{U}_{qq}^{-1}\mathbf{b}_q.$$

§ 4. Modele reduse

Studiem efectul ipotezelor $\mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$, $\mathbf{C}'\boldsymbol{\beta} = \mathbf{0}$ şi $\boldsymbol{\beta}_q = \mathbf{0}$ asupra modelului $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

i) $\mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$. În estimarea lui $\boldsymbol{\beta}$ supusă la condiţia $\mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$, spunem că avem de a face cu modelul $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ asupra căruia s-a impus constrîngerea $\mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$. Avem

DEFINIŢIA 14. Modelul $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ fără constrîngeri, poartă numele de model maxim şi modelul $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ cu $\mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$ poartă numele de model redus,

Exemplul 1. Dacă modelul maxim este

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

şi ipoteza este $H: \beta_1 = \beta_2$, modelul redus este

$$Y_i = \beta_0 + \beta_1(X_{i1} + X_{i2}) + \beta_3 X_{i3} + \varepsilon_i.$$

Cercetăm semnificaţia statisticilor Q şi SSE + Q în funcţie de SS asociate cu modelele: maxim şi redus. În acest scop introducem termenii:

reducere (maxim) şi rezidual (maxim)

pentru SS redusă şi respectiv reziduală după aplicarea modelului maxim:

redus(maxim) = SSR şi rezidual(maxim) = SSE.

Similar (din propoziţia 5)

(92) $$\text{SSE} + Q = \text{rezidual (redus)}.$$

Deci

$$Q = \text{SSE} + Q - \text{SSE}$$

(93) $$= \text{rezidual (redus)} - \text{rezidual (maxim)}$$

Scanned with OKEN Scanner

și

$$Q = \mathbf{Y}'\mathbf{Y} - \text{SSE} - [\mathbf{Y}'\mathbf{Y} - (\text{SSE} + Q)] =$$

$$= \text{SSR} - [\mathbf{Y}'\mathbf{Y} - (\text{SSE} + Q)]$$

(94)
$$= \text{reducere (maxim)} - [\mathbf{Y}'\mathbf{Y} - (\text{SSE} + Q)].$$

Din compararea lui (94) cu (93), sîntem tentaţi să tragem concluzia că

$$\mathbf{Y}'\mathbf{Y} - (\text{SSE} + Q) = \text{redus (redus)},$$

reducerea în SS datorată ajustării modelului redus. Aceasta se datoreşte faptului că :

$$\text{SSE} + Q = \text{rezidual (redus)}.$$

Totuşi, vom arăta că numai în cazuri speciale

$$\mathbf{Y}'\mathbf{Y} - (\text{SSE} + Q)$$

este reducerea în SS datorată ajustării modelului redus.

Mai întîi arătăm că $\mathbf{Y}'\mathbf{Y} - (\text{SSE} + Q)$ nu este în general o SS, deoarece ea poate fi negativă. Astfel în

$$\mathbf{Y}'\mathbf{Y} - (\text{SSS} + Q) = \text{SSR} - Q =$$

(95)
$$= \mathbf{b}'\mathbf{X}'\mathbf{Y} - (\mathbf{C}'\mathbf{b} - \mathbf{m})' [\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}(\mathbf{C}'\mathbf{b} - \mathbf{m})$$

al doilea termen este o formă semipozitiv definită. Prin urmare nu este niciodată negativ şi dacă unul sau mai multe elemente ale lui $\mathbf{m}$ sînt suficient de mari, termenul va depăşi $\mathbf{b}'\mathbf{X}'\mathbf{Y}$ şi (95) va fi negativ, Deci $\mathbf{Y}'\mathbf{Y} - (\text{SSE} + Q)$ nu este o SS.

Motivul pentru care $\mathbf{Y}'\mathbf{Y} - (\text{SSE} + Q)$ nu este necesar o reducere în SS datorită ajustării modelului redus este că $\mathbf{Y}'\mathbf{Y}$ nu este întotdeauna suma totală a pătratelor pentru modelul redus.

Exemplul 2. Dacă modelul maxim este

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

şi ipoteza este $\beta_1 = \beta_2 + \gamma$, atunci modelul redus va fi

$$Y_i = \beta_0 + (\beta_2 + \gamma) X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

sau

(96)
$$Y_i - \gamma X_{i1} = \beta_0 + \beta_2(X_{i1} + X_{i2}) + \varepsilon_i.$$

Suma totală a pătratelor pentru acest model redus este

$$(\mathbf{Y} - \gamma\mathbf{X}_1)' (\mathbf{Y} - \gamma\mathbf{X}_1)$$

Scanned with OKEN Scanner

şi nu $\mathbf{Y'Y}$ şi astfel

$$\mathbf{Y'Y} - (\mathrm{SSE} + Q)$$

nu este SS redusă.

Mai mult (96) nu este singurul model redus, deoarece ipoteza $\beta_1 = \beta_2 + \gamma$ poate fi folosită pentru a da modelul

$$Y_i = \beta_0 + \beta_1 X_{i1} + (\beta_1 - \gamma) X_{i2} + \varepsilon_i$$

sau

$$Y_i + \gamma X_{i2} = \beta_0 + \beta_1 (X_{i1} + X_{i2}) + \varepsilon_i. \tag{97}$$

Acum suma totală a pătratelor va fi

$$(\mathbf{Y} + \gamma \mathbf{X}_2)' (\mathbf{Y} + \gamma \mathbf{X}_2).$$

Astfel în acest caz există două modele (96) şi (97) şi ele nu sînt identice; nici unul dintre ele nu are suma totală a pătratelor egală cu $\mathbf{Y'Y}$. Prin urmare

$$\mathbf{Y'Y} - (\mathrm{SSE} + Q)$$

nu este reducerea în SS datorită ajustării modelului redus. Din existenţa lui (96) şi (97) nu există un singur model redus. În ciuda acestui fapt, $\mathrm{SSE} + Q$ este SS reziduală pentru toate modelele reduse posibile, SS totale şi reducerile în SS diferă de la un model la altul, însă SS reziduale sînt aceleaşi.

Situaţia descrisă anterior este adevărată pentru ipoteza $\mathbf{C'\beta} = \mathbf{m}$. Presupunem că $\mathbf{L'}$ este astfel încît

$$\mathbf{R} = \begin{pmatrix} \mathbf{C'} \\ \mathbf{L'} \end{pmatrix} \text{ are rang maxim}$$

şi $\mathbf{R}^{-1} = (\mathbf{P}\ \mathbf{S})$ este inversa sa. Atunci modelul $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ poate fi scris ca

$$\mathbf{Y} = \mathbf{X}\mathbf{R}^{-1}\mathbf{R}\boldsymbol{\beta} + \boldsymbol{\varepsilon} =$$

$$= \mathbf{X}(\mathbf{P}\ \mathbf{S}) \begin{pmatrix} \mathbf{C'\beta} \\ \mathbf{L'\beta} \end{pmatrix} + \boldsymbol{\varepsilon} =$$

$$= \mathbf{XPm} + \mathbf{XSL'\beta} + \boldsymbol{\varepsilon}$$

sau

$$\mathbf{Y} - \mathbf{XPm} = \mathbf{XSL'\beta} + \boldsymbol{\varepsilon}. \tag{98}$$

Acesta este un model în elementele lui $\mathbf{L'\beta}$, care reprezintă $r - s$ funcţii liniare de elementele lui $\boldsymbol{\beta}$. Însă deoarece $\mathbf{L'}$ este arbitrară, o alegem pentru a face $\mathbf{R}$ nesingulară; modelul (98) nu este unic. Cu toate

acestea, SS reziduală după ajustarea oricărui model (98) este SSE + Q și valoarea corespunzătoare a estimatorului parametrului $\boldsymbol{\beta}$ este $\tilde{\mathbf{b}}$ dat de (74).

ii) $\mathbf{C}'\boldsymbol{\beta} = \mathbf{0}$. Unul din cazurile în care $\mathbf{Y}'\mathbf{Y} - (\text{SSE} + Q)$ este reducerea în SS datorită ajustării modelului redus se obține când $\mathbf{m} = \mathbf{0}$. Atunci (98) devine

$$\mathbf{Y} = \mathbf{X}\mathbf{S}\mathbf{L}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

și astfel suma totală a pătratelor pentru modelul redus este $\mathbf{Y}'\mathbf{Y}$, aceeași pentru modelul maxim. Deci în acest caz

$$\mathbf{Y}'\mathbf{Y} - (\text{SSE} + Q) = \text{reducerea (redus)}. \tag{99}$$

Aceasta este o SS, adică este semipozitiv definită; astfel din (95) punînd $\mathbf{m} = \mathbf{0}$, obținem:

$$\begin{aligned}\mathbf{Y}'\mathbf{Y} - (\text{SSE} + Q) &= \mathbf{b}'\mathbf{X}'\mathbf{Y} - \mathbf{b}'\mathbf{C}[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]\mathbf{C}'\mathbf{b} = \\ &= \mathbf{Y}'\{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}\mathbf{Y}.\end{aligned} \tag{100}$$

Cum matricea din { } este idempotentă ea este semipozitiv definită. Prin urmare așa este și $\mathbf{Y}'\mathbf{Y} - (\text{SSE} + Q)$; adică este o sumă de patrate.

Din (99) avem:

$$Q = \mathbf{Y}'\mathbf{Y} - \text{SSE} - \text{reducerea (redus)}.$$

Însă

$$\mathbf{Y}'\mathbf{Y} - \text{SSE} = \text{SSR} = \text{reducere (maxim)}$$

și astfel

$$Q = \text{reducere (maxim)} - \text{reducere (redus)}.$$

Prin urmare, deoarece singura diferență dintre modelele maxim și redus este ipoteza, are loc.

DEFINIȚIA 15. Q este reducerea în SS datorită ipotezei puse.

Din tabelul 1, partiționînd

$$\text{SSR} = Q + (\text{SSR} - Q)$$

obținem

Tabelul 4

Analiza dispersiei pentru verificarea ipotezei $\mathbf{C}'\boldsymbol{\beta} = \mathbf{0}$

Suma variației	Gr. de libertate	SS
Regresia (modelul maxim)	r	SSR
Ipoteza	s	Q
Modelul redus	$r - s$	SSR$-Q$
Eroarea reziduală	$n - r$	SSE
Total	n	SST

Scanned with OKEN Scanner

COROLARUL 9. *Avem*

$$Q/\sigma^2 \sim \chi^2\{s, \boldsymbol{\beta}'\mathbf{C}[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}\mathbf{C}'\boldsymbol{\beta}/2\sigma^2\}.$$

Demonstrația rezultă din (69) pentru $\mathbf{m} = \mathbf{0}$.

TEOREMA 8. *Avem*

$$(\mathrm{SSR} - Q)/\sigma^2 \sim \chi^2\{r - s, \boldsymbol{\beta}'[\mathbf{X}'\mathbf{X} - \mathbf{C}\{\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'\}^{-1}\mathbf{C}']\boldsymbol{\beta}/2\sigma^2\}$$

și este independentă de SSE/σ^2

Demonstrație. Deoarece

$$(\mathbf{Y}'\mathbf{Y} - \mathrm{SSE})/\sigma^2 \sim \chi^2\{r, \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}/2\sigma^2\}$$

teorema este demonstrată ținînd seama de problema 14.

În funcție de (98), parametrul de necentricitate al repartiției statisticii $\mathrm{SSR} - Q$ este

$$\boldsymbol{\beta}'\mathbf{L}(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S})\mathbf{L}'\boldsymbol{\beta}/2\,\sigma^2.$$

Deci în ipoteza nulă, acest parametru de necentricitate este zero cînd $\mathbf{L}'\boldsymbol{\beta} = \mathbf{0}$. Astfel

$$\mathrm{SSR} - Q$$

poate fi folosită ca statistică într-un test F pentru cazul $\mathbf{L}'\boldsymbol{\beta} = \mathbf{0}$ sub ipoteza nulă $\mathbf{C}'\boldsymbol{\beta} = \mathbf{0}$.

Cu acestea avem următoarele statistici F

(101) $\dfrac{\mathrm{SSR}/r}{\mathrm{SSE}/(n-r)}$ pentru verificarea modelului maxim

(102) $\dfrac{Q/s}{\mathrm{SSE}/(n-r)}$ pentru verificarea ipotezei $\mathbf{C}'\boldsymbol{\beta} = \mathbf{0}$

și în ipoteza nulă,

(103) $\dfrac{(\mathrm{SSR} - Q)/(r-s)}{\mathrm{SSE}/(n-r)}$ pentru verificarea subipotezei $\mathbf{L}'\boldsymbol{\beta} = \mathbf{0}$.

iii) $\boldsymbol{\beta}_q = \mathbf{0}$. Cel mai folosit caz al modelului redus cînd $\mathbf{m} = \mathbf{0}$ este cînd $\mathbf{C}' = (\mathbf{I}_q \quad \mathbf{0})$ pentru $q \leqslant k$. Atunci ipoteza nulă $\mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$ este

$$\boldsymbol{\beta}_q = \mathbf{0}, \text{ unde } \boldsymbol{\beta}_q' = (\beta_0, \beta_1 \ldots \beta_{q-1}).$$

Această situație am prezentat-o atunci cînd am obținut relația (88)

$$F(H) = Q/q\hat{\sigma}^2 \text{ cu } Q = \mathbf{b}_q'\mathbf{a}_{qq}^{-1}\mathbf{b}_q.$$

Deci un caz special al tabelului 4 este analiza dispersiei pentru verificarea ipotezei $H : \boldsymbol{\beta}_q = \mathbf{0}$ sintetizată în tabelul 5.

Tabelul 5

Analiza dispersiei pentru verificarea ipotezei $\boldsymbol{\beta}_q = \mathbf{0}$

Sursa variației	Gr. de libertate	SS
Modelul maxim ($\boldsymbol{\beta}$)	r	$\mathrm{SSR} = \mathbf{b}'\mathbf{X}'\mathbf{Y}$
Ipoteză : $\boldsymbol{\beta}_q = \mathbf{0}$	q	$Q = \mathbf{b}_q'\mathscr{U}_{qq}^{-1}\mathbf{b}_q$
Modelul redus ($\boldsymbol{\beta}_q$)	$r - q$	$\mathrm{SSR} - Q$
Eroarea reziduală	$n - r$	$\mathrm{SSE} = \mathrm{SST} - \mathrm{SSR}$
Total	n	$\mathrm{SST} = \mathbf{Y}'\mathbf{Y}$

$\mathrm{SSR} - Q$ poate fi exprimat ca $\tilde{\mathbf{b}}_p'\mathbf{X}_p'\mathbf{X}_p\tilde{\mathbf{b}}_p$, estimatorul $\tilde{\mathbf{b}}_p$ fiind obținut din (74) ca

$$\tilde{\mathbf{b}}_q = \mathbf{b}_p - \mathscr{U}_{pq}\mathscr{U}_{qq}^{-1}\mathbf{b}_q. \tag{104}$$

§ 5. Examinarea rezidurilor

DEFINIȚIA 16. Statistica

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}\mathbf{b} \tag{105}$$

poartă numele de vectorul rezidurilor.

Componentele $\hat{\varepsilon}_i$ ale acestui vector sînt reprezentate grafic și cercetate pentru a vedea dacă presupunerile legate de model sînt sau nu corecte. Putem să scriem :

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}\mathbf{b} = \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{K}\mathbf{Y}.$$

Avem

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = \mathbf{1}'\hat{\boldsymbol{\varepsilon}} = \mathbf{1}'\mathbf{K}\mathbf{Y}$$

și cum $\mathbf{1}'\mathbf{K} = \mathbf{0}'$ obținem

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0.$$

Din (39), găsim

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \mathbf{Y}'\mathbf{K}'\mathbf{K}\mathbf{Y} = \mathbf{Y}'\mathbf{K}\mathbf{Y} = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathrm{SSE}.$$

Scanned with OKEN Scanner

Avem

$$M(\hat{\boldsymbol{\varepsilon}}) = M(\mathbf{KY}) = \mathbf{KX\beta} = \mathbf{0} \text{ (din (38))}$$

$$\mathscr{C}(\hat{\boldsymbol{\varepsilon}}) = \mathscr{C}(\mathbf{KY}) = \mathbf{K}^2\sigma^2 = \mathbf{K}\sigma^2 \text{ (din (37))}$$

Astfel am demonstrat

TEOREMA 9. i) *Suma reziduriloг este zero,*

ii) SS *reziduriloг este* SSE,

iii) $M(\hat{\boldsymbol{\varepsilon}}) = 0$ *şi*

iv) $\mathscr{C}(\hat{\boldsymbol{\varepsilon}}) = \mathbf{K}\sigma^2$.

De asemenea are loc

TEOREMA 10 i) $\operatorname{cov}(\hat{\boldsymbol{\varepsilon}}, \mathbf{Y}) = \mathbf{K}\sigma^2$, $\operatorname{cov}(\hat{\boldsymbol{\varepsilon}}, \hat{\mathbf{Y}}) = \mathbf{0}_{n\times n}$

ii) $\operatorname{cov}(\hat{\boldsymbol{\varepsilon}}, \mathbf{b}) = \mathbf{0}_{n\times(k+1)}$, $\operatorname{cov}(\boldsymbol{\varepsilon}, \mathbf{b}) = \sigma^2\mathbf{I}$

iii) $\sum_{i=1}^{n} \hat{\varepsilon}_i Y_i = \text{SSE}$, $\sum_{i=1}^{n} \hat{\varepsilon}_i \hat{Y}_i = 0$.

Aceste rezultate au loc pentru orice model cu intercepţie.

TEOREMA 11. *Dacă* $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, atunci

$$\hat{\boldsymbol{\varepsilon}} \sim \mathrm{N}(\mathbf{0}, \mathbf{K}\sigma^2).$$

Se reprezintă grafic valorile $\hat{\varepsilon}_i$ pentru a vedea dacă presupunerea $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$ este sau nu corectă. Făcînd acest lucru se ignoră faptul că deoarece $\mathscr{C}(\hat{\boldsymbol{\varepsilon}}) = \mathbf{K}\sigma^2$, $\hat{\varepsilon}_i$ sînt corelate.

Alte procedee grafice care pot pune în evidenţă defecţiuni în presupuneri, reprezintă rezidurile faţă de $\widehat{M(Y)}$ şi faţă de valorile observate ale lui X. Ultima reprezentare se foloseşte cînd X-şii reprezintă timpi şi chiar dacă acesta nu este cazul, dacă timpul a fost un factor implicat în colectarea datelor. Se pot trage două concluzii din acest fel de reprezentări : că dispersia termenilor eroare nu este constantă sau că sînt necesari termeni adiţionali în model.

DEFINIŢIA 17 (Larsen, McCleary (1972)). Statistica

$$\hat{\boldsymbol{\varepsilon}}_i^* = \hat{\boldsymbol{\varepsilon}} + \mathbf{X}_i b_i \tag{106}$$

unde $\hat{\boldsymbol{\varepsilon}}$ este dat de (105) şi $\mathbf{X}_i$ este coloana i a matricii $\mathbf{X}$, poartă numele de vectorul parţial al rezidurilor.

Ţinînd seama de (105), (106) se scrie

$$\hat{\boldsymbol{\varepsilon}}_i^* = \mathbf{Y} - \sum_{j\neq i} \mathbf{X}_j b_j. \tag{106'}$$

Reprezentarea grafică a vectorului parţial al rezidurilor stabileşte

— importanţa (în termeni de putere de predicţie pentru variabila dependentă) variabilei independente i în prezenţa celorlalte variabile independente şi

— importanţa neliniarităţii.

Scanned with OKEN Scanner

Prezintă importanţă verificarea ipotezei conform căreia rezidurile au un caracter întîmplător. Considerînd şirul rezidurilor obţinem grupuri de reziduri unele avînd semnul plus iar altele semnul minus, deci iteraţii.

Notînd cu n_1 = numărul semnelor +, cu n_2 = numărul semnelor minus şi cu u numărul de iteraţii, din teoremele XIX.4 şi 5 avem

$$M(u) = \frac{2n_1n_2 + n}{n} \tag{107}$$

$$D^2(u) = \frac{2n_1n_2(2n_1n_2 - 1)}{(n-1)\,n^2} \tag{108}$$

şi aproximativ

$$Z = \frac{u - M(u) + \frac{1}{2}}{D(u)} \sim N(0, 1), \tag{109}$$

$\frac{1}{2}$ fiind corecţia pentru continuitate.

Folosind aproximaţia (109) se verifică ipoteza conform căreia rezidurile sînt aleatoare (întîmplătoare).

De asemenea prezintă importanţă eliminarea acelor reziduri, care în valoare absolută, sînt mult mai mari decît restul celorlalte (în general numindu-se reziduri nesemnificative) şi totodată cercetarea cauzelor ce au generat astfel de reziduri.

Pentru detalii privind caracterul întîmplător şi eliminarea rezidurilor nesemnificative se poate vedea Craiu, Bad (1979).

Scanned with OKEN Scanner

Capitolul XXXII

MODELE CARE NU SÎNT DE RANG MAXIM

§ 1. Modelul. Presupuneri

Ne ocupăm de modelul

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

unde $\mathbf{Y}$ este vectorul observaţiilor $(n \times 1)$
$\boldsymbol{\beta}$ este vectorul parametrilor $(p \times 1)$ şi
$\mathbf{X}$ o matrice $n \times p$ de valori cunoscute (în cele mai multe cazuri 0 sau 1), unde $\mathbf{X}$ nu are rang maxim coloană şi
$\boldsymbol{\varepsilon}$ este vectorul erorilor (aleator).
$\boldsymbol{\varepsilon}$ poate fi definit ca :

$$\boldsymbol{\varepsilon} = \mathbf{Y} - M(\mathbf{Y}),$$

astfel încît

$$M(\boldsymbol{\varepsilon}) = 0 \quad \text{şi} \quad M(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}.$$

Se presupune că fiecare componentă a vectorului $\boldsymbol{\varepsilon}$ are dispersia σ^2 şi covarianţa cu orice altă componentă zero, adică

$$\mathscr{C}(\boldsymbol{\varepsilon}) = M(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}_n.$$

Astfel

$$\boldsymbol{\varepsilon} \sim (0, \sigma^2\mathbf{I}_n) \quad \text{şi} \quad \mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n).$$

§ 2. Estimarea parametrilor

2.1. Ecuaţiile normale

Ca şi în capitolul XXXI ecuaţiile normale sînt

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}. \tag{1}$$

Scanned with OKEN Scanner

Ori de cîte ori $\mathbf{X'X}$ nu este de rang maxim, ecuaţiile (1) nu au o singură soluţie $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$ ca în capitolul XXXI. Sînt posibile mai multe soluţii. Pentru a pune în evidenţă acest fapt scriem ecuaţiile normale ca

$$\mathbf{X'Xb^0} = \mathbf{X'Y} \tag{2}$$

folosind simbolul $\mathbf{b^0}$ pentru a distinge soluţiile ecuaţiilor (2) de unica soluţie care există cînd $\mathbf{X'X}$ are rang maxim. Pentru a obţine una dintre ele găsim orice inversă generalizată $\mathbf{G}$ a matricii $\mathbf{X'X}$ şi scriem soluţia corespunzătoare ca

$$\mathbf{b^0} = \mathbf{GX'Y}. \tag{3}$$

Notaţia $\mathbf{b^0}$ pentru soluţia ecuaţiilor normale (2), accentuează că $\mathbf{b^0}$ obţinută rezolvînd (2) este numai o soluţie a ecuaţiei şi nu un estimator pentru $\boldsymbol{\beta}$. În modelele liniare care nu sînt de rang maxim, ceea ce am obţinut ca soluţie a ecuaţiilor normale este o soluţie şi nimic mai mult.

Este adevărat că $\mathbf{b^0}$ este un estimator, însă nu pentru $\boldsymbol{\beta}$ şi expresia acestei estimaţii depinde de ce inversă generalizată pentru $\mathbf{X'X}$ a fost folosită în obţinerea lui $\mathbf{b^0}$.

2.2. Proprietăţile soluţiei ecuaţiilor normale

TEOREMA 1. *Avem*

$$M(\mathbf{b^0}) = \mathbf{H}\boldsymbol{\beta} \tag{4}$$

şi

$$\mathscr{C}(\mathbf{b^0}) = \mathbf{GX'XG'}\,\sigma^2 \tag{5}$$

unde $\mathbf{H} = \mathbf{GX'X}$.

Demonstraţie. Din (3) obţinem

$$M(\mathbf{b^0}) = \mathbf{GX'}M(\mathbf{Y}) = \mathbf{GX'X}\boldsymbol{\beta} = \mathbf{H}\boldsymbol{\beta}.$$

Deci $\mathbf{b^0}$ este un estimator nedeplasat pentru $\mathbf{H}\boldsymbol{\beta}$, însă nu pentru $\boldsymbol{\beta}$.

Din (3) găsim

$$\mathscr{C}(b^0) = \mathbf{GX'}\mathscr{C}(\mathbf{Y})\mathbf{XG'} =$$

$$= \mathbf{GX'XG'}\,\sigma^2.$$

Observaţie. Cu o alegere convenabilă a matricii $\mathbf{G}$, (5) poate fi exprimată ca $\mathbf{G}\sigma^2$. Deoarece $\mathbf{X'X}$ este simetrică, există o matrice de permutare $\mathbf{P}$ ortogonală astfel încît

$$\mathbf{P'X'XP} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}'_{12} & \mathbf{A}_{22} \end{pmatrix}$$

unde $\mathbf{A}_{11}$ este pătrată, de rang maxim, egal cu rangul lui $\mathbf{X'X}$. Atunci

$$\mathbf{G} = \mathbf{P}\begin{pmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{P'}$$

Scanned with OKEN Scanner

este inversa generalizată simetrică a lui $\mathbf{X'X}$ cu $\mathbf{GX'XG'} = \mathbf{G}$. Deci cu acest $\mathbf{G}$,

$$\mathscr{C}(\mathbf{b}^0) = \mathbf{G}\sigma^2. \quad .$$

Totuși nu avem nici un avantaj din această formă.

Corespunzător vectorului observațiilor $\mathbf{Y}$, avem

$$\widehat{M(\mathbf{Y})} \equiv \hat{\mathbf{Y}} = \mathbf{Xb}^0 = \mathbf{XGX'Y}. \tag{6}$$

PROPOZIȚIA 1. *Vectorul* $\widehat{M(\mathbf{Y})}$ *este invariant la alegerea inversei generalizate a matricei* $\mathbf{X'X}$ *care este folosită pentru* $\mathbf{G}$.

Demonstrația rezultă din faptul că $\mathbf{XGX'}$ este invariantă.

Observație. Aceasta arată că indiferent de soluția ecuației normale folosită pentru $\mathbf{b}^0$ vectorul $\hat{\mathbf{Y}} = \mathbf{XGX'Y}$ este întotdeauna același. Cu alte cuvinte putem alege o soluție a ecuațiilor normale (în orice mod dorim), o notăm $\mathbf{b}^0$, și indiferent ce soluție este, $\hat{\mathbf{Y}} = \mathbf{Xb}^0$ va fi valoarea corectă a lui $\hat{\mathbf{Y}}$.

2.3. Sume de pătrate. Estimarea lui σ^2

DEFINIȚIA 1. Statistica

$$\mathrm{SSE} = (\mathbf{Y} - \mathbf{Xb}^0)'(\mathbf{Y} - \mathbf{Xb}^0)$$

poartă numele de SS erorilor.

TEOREMA 2. *Avem*

(7) i) $\mathrm{SSE} = \mathbf{Y'}(\mathbf{I} - \mathbf{XGX'})\mathbf{Y}$

(8) ii) $\mathrm{SSE} = \mathbf{Y'Y} - \mathbf{b}^{0\prime}\mathbf{X'Y}$

iii) SSE *este invariantă la alegerea soluției ecuației normale care este folosită pentru* $\mathbf{b}^0$

iv) $\mathbf{M}(\mathrm{SSE}) = [n - r(\mathbf{X})]\sigma^2$ *sau*

(9) iv') $\hat{\sigma}^2 = \mathrm{SSE}/[n - r(\mathbf{X})]$ *este un estimator nedeplasat pentru* σ^2.

Demonstrație. Avem

$$\mathrm{SSE} = \mathbf{Y'}(\mathbf{I} - \mathbf{XGX'})(\mathbf{I} - \mathbf{XGX'})\mathbf{Y},$$

ceea ce demonstrează i).

Cum

$$\mathrm{SSE} = \mathbf{Y'Y} - \mathbf{Y'XGX'Y},$$

obținem ii).

Deoarece $\mathbf{Y} \sim (\mathbf{X\beta}, \sigma^2\mathbf{I})$,

$$M(\mathrm{SSE}) = M[\mathbf{Y'}(\mathbf{I} - \mathbf{XGX'})\mathbf{Y}] = \mathrm{Ur}[(\mathbf{I} - \mathbf{XGX'})\mathbf{I}\sigma^2] + \\ + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{XGX'})\mathbf{X}\boldsymbol{\beta}$$

Scanned with OKEN Scanner

şi ţinînd seama de problema 6

$$M(\mathrm{SSE}) = \sigma^2 r(\mathbf{I} - \mathbf{XGX}') = [n - r(\mathbf{X})]\sigma^2,$$

ceea ce demonstrează iv) şi iv').

În cazul modelelor care nu sînt de rang maxim, pentru partiţia sumei totale a pătratelor avem (corespunzător la (XXXI.44), XXXI.44') şi (XXXI.47))

$$(10) \qquad \begin{cases} \mathrm{SSR} = \mathbf{Y}'\mathbf{XGX}'\mathbf{Y} \\ \mathrm{SSE} = \mathbf{Y}'(\mathbf{I} - \mathbf{XGX}')\mathbf{Y} \\ \mathrm{SST} = \mathbf{Y}'\mathbf{Y} \end{cases}$$

în care

$$(11) \qquad \mathrm{SSR} = \mathrm{SST} - \mathrm{SSE} = \mathbf{Y}'\mathbf{X}'\mathbf{GXY} = \mathbf{b}^{0\prime}\mathbf{X}'\mathbf{Y}$$

este SS atribuită ajustării modelului $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$;

$$(12) \qquad \begin{cases} \mathrm{SSM} = n\bar{Y}^2 = \mathbf{Y}'n^{-1}\mathbf{11}'\mathbf{Y} \\ \mathrm{SSR}_m = \mathbf{Y}'(\mathbf{XGX}' - n^{-1}\mathbf{11}')\mathbf{Y} \\ \mathrm{SSE} = \mathbf{Y}'(\mathbf{I} - \mathbf{XGX}')\mathbf{Y} \\ \mathrm{SST} = \mathbf{Y}'\mathbf{Y} \end{cases}$$

în care: SSM este SS datorită ajustării mediei generale şi

$$(13) \qquad \mathrm{SSR}_m = \mathrm{SSR} - \mathrm{SSM} = \mathrm{SSR} - n\bar{Y}^2$$

este SS datorită ajustării modelului, corectat pentru medie şi respectiv

$$(14) \qquad \begin{cases} \mathrm{SSR}_m = \mathbf{Y}'(\mathbf{XGX}' - n^{-1}\mathbf{11}')\mathbf{Y} \\ \mathrm{SSE} = \mathbf{Y}'(\mathbf{I} - \mathbf{XGX}')\mathbf{Y} \\ \mathrm{SST}_m = \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2 \end{cases}$$

în care

$$(15) \qquad \mathrm{SST}_m = \mathrm{SST} - \mathrm{SSM} = \sum Y_i^2 - n\bar{Y}^2.$$

Ne vom referi la pătratul coeficientului de corelaţie dintre valorile observate Y şi elementele corespunzătoare ale lui $\hat{\mathbf{Y}}$ ca la coeficientul determinării.

$$(16) \qquad R^2 = \textit{coeficientul determinării} = \frac{[\sum(Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})]^2}{\sum(Y_i - \bar{Y})^2 \sum(Y_i - \bar{\hat{Y}})^2}$$

Propoziţia 2. *Avem*

$$(17) \qquad R^2 = \frac{(\mathrm{SSR}_m)^2}{\mathrm{SST}_m(\mathrm{SSR}_m)} = \frac{\mathrm{SSR}_m}{\mathrm{SST}_m}.$$

Scanned with OKEN Scanner

Demonstrație. Din problema 6

$$\mathbf{X'XGX'} = \mathbf{X'}$$

și deoarece $\mathbf{1'}$ este prima linie a lui $\mathbf{X'}$

$$\mathbf{1'XGX'} = \mathbf{1'}. \tag{18}$$

Avem $\bar{\hat{Y}} = \bar{Y}$, ceea ce demonstrează propoziția.

2.4. Repartiția soluției ecuațiilor normale. Repartiții ale sumelor de pătrate

TEOREMA 3. *Dacă* $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$ *atunci*

i) $\mathbf{b^0} = \mathbf{GX'Y} \sim N(\mathbf{H\beta}, \mathbf{GX'XG'}\sigma^2)$

ii) $\mathbf{b^0}$ *și* $\hat{\sigma}^2$ *sînt independente.*

Demonstrație. Folosind problema 12 pentru

$$\mathbf{b^0} = \mathbf{GX'Y} \text{ și } SSE = \mathbf{Y'(I - XGX')Y}$$

se vede că $\mathbf{GX'I}\sigma^2(\mathbf{I - XGX'}) = \mathbf{G(X' - X'XGX')}\sigma^2 = \mathbf{0}$ (deoarece $\mathbf{X'} = \mathbf{X'XGX}$ (problema 6). Prin urmare $\mathbf{b^0}$ și $\hat{\sigma}^2$ sînt independente.

TEOREMA 4. i) $SSE/\sigma^2 \sim \chi^2(n - r)$

ii) $SSR/\sigma^2 \sim \chi^2(r, \boldsymbol{\beta'X'X\beta}/2\sigma^2)$ *și independentă de* SSE

iii) $SSM/\sigma^2 \sim \chi^2[1, (\mathbf{1'X\beta})^2/2n\sigma^2]$; SSM *este independentă de* SSE

iv) $SSR_m/\sigma^2 \sim \chi^2[r - 1, \boldsymbol{\beta'X'}(\mathbf{I} - n^{-1}\mathbf{11'})\mathbf{X\beta}/2\sigma^2]$; SSR_m *este independentă de* SSM și SSE.

v) $SSR_m/\sigma^2 \sim \chi^2[r - 1, \boldsymbol{\beta}^{*\prime}\mathcal{A}'\mathcal{A}\boldsymbol{\beta}^*/2\sigma^2]$.

Demonstrație. i) Avem

$$SSE/\sigma^2 = \mathbf{Y'(I - XGX')Y}/\sigma^2$$

și aplicînd problema 11, găsim

$$\mathbf{I}\sigma^2(\mathbf{I - XGX'})/\sigma^2 = \mathbf{I - XGX'}$$

matrice ce este indempotentă. Prin urmare din problema 11

$$SSE/\sigma^2 \sim \chi^2[r(\mathbf{I - XGX'}), \boldsymbol{\beta'X'}(\mathbf{I - XGX'})\mathbf{X\beta}/2\sigma^2],$$

care din proprietățile lui $\mathbf{XGX'}$ și cu $r(X) = r$ se reduce la i).

ii) Din (11)

$$SSR = \mathbf{Y'XGX'Y}$$

în care $\mathbf{XGX'}$ este indempotentă și

$$\mathbf{XGX'(I - XGX')} = \mathbf{0}.$$

Scanned with OKEN Scanner

Prin urmare din problemele 13 şi 11 SSR/σ^2 este independentă de SSE cu

$$SSR/\sigma^2 \sim \chi^2[r(\mathbf{XGX}'), \boldsymbol{\beta}'\mathbf{X}'\mathbf{XGX}'\mathbf{X}\boldsymbol{\beta}/2\sigma^2],$$

ceea ce demonstrează ii).

iii) Avem

$$SSM/\sigma^2 = \mathbf{Y}'n^{-1}\mathbf{11}'\mathbf{Y}/\sigma^2$$

unde $n^{-1}\mathbf{11}'$ este idempotentă. Din (18)

$$n^{-1}\mathbf{11}'\mathbf{X}'\mathbf{GX}' = n^{-1}\mathbf{11}' \tag{19}$$

astfel încît

$$n^{-1}\mathbf{11}'(\mathbf{I} - \mathbf{XGX}') = \mathbf{0}.$$

Deci SSM este independentă de SSE şi

$$SSM/\sigma^2 \sim \chi^2[r(n^{-1}\mathbf{11}'), \boldsymbol{\beta}'\mathbf{X}'n^{-1}\mathbf{11}'\mathbf{X}\boldsymbol{\beta}/2\sigma^2] \tag{20}$$

ceea ce demonstrează iii).

iv) Raţionamentul este acelaşi pentru

$$SSR_m = \mathbf{Y}'(\mathbf{XGX}' - n^{-1}\mathbf{11}')\mathbf{Y}.$$

Din (19), $\mathbf{XGX}' - n^{-1}\mathbf{11}'$ este idempotentă şi

$$(\mathbf{XGX}' - n^{-1}\mathbf{11}')n^{-1}\mathbf{11}' = (\mathbf{XGX}' - n^{-1}\mathbf{11}')(\mathbf{I} - \mathbf{XGX}') = \mathbf{0}.$$

Deci din problemele 11 şi 13, SSR_m este independentă de SSM şi SSE, şi

$$SSR_m/\sigma^2 \sim \chi^2[r(\mathbf{XGX}' - n^{-1}\mathbf{11}'), \boldsymbol{\beta}'\mathbf{X}'(\mathbf{XGX}' - n^{-1}\mathbf{11}')\mathbf{X}\boldsymbol{\beta}/2\sigma^2], \tag{21}$$

ceea ce demonstrează iv).

v) Ori de cîte ori $\mathbf{X}$ are ca primă coloană pe $\mathbf{1}$, putem scrie $\mathbf{X} = (\mathbf{1}\ \mathbf{X}_1)$ şi atunci

$$\mathbf{X}'(\mathbf{I} - n^{-1}\mathbf{11}')\mathbf{X} = \begin{pmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \mathbf{X}_1'(\mathbf{I} - n^{-1}\mathbf{11}')\mathbf{X}_1 \end{pmatrix} = \begin{pmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \mathscr{A}'\mathscr{A} \end{pmatrix}, \tag{22}$$

unde $\mathscr{A}'\mathscr{A}$ reprezintă acelaşi fel de matrice ca aceea din capitolul XXXI şi anume SS şi produse de abateri ale elementelor coloanelor (altele decît prima coloană) matricii $\mathbf{X}$ de la mediile lor. Simbolic este mai simplă decît forma sa echivalentă

$$\mathbf{X}_1'(\mathbf{I} - n^{-1}\mathbf{11}')\mathbf{X}_1$$

Scanned with OKEN Scanner

însă oferă un avantaj mic, spre deosebire de modelul de rang maxim unde este avantajos.

La fel ca în cazul modelului de rang maxim, punînd

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}^* \end{pmatrix}, \tag{23}$$

cu β_0 reprezentînd o medie generală, din (21) și (22) avem

$$\mathrm{SSR}_m/\sigma^2 \sim \chi^2[r-1, \boldsymbol{\beta}^{*\prime}\mathbf{X}_1'(\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}')\mathbf{X}_1\boldsymbol{\beta}^*/2\sigma^2], \tag{24}$$

ceea ce demonstrează v). Teorema este demonstrată.

COROLARUL 1. *Avem*

$$F(R) = \frac{\mathrm{SSR}/r}{\mathrm{SSE}/(n-r)} \sim F(r, n-r, \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}/2\sigma^2) \tag{25}$$

$$F(M) = \frac{\mathrm{SSM}/1}{\mathrm{SSE}/(n-r)} \sim F[1, n-r, (\mathbf{1}'\mathbf{X}\boldsymbol{\beta})^2/2n\sigma^2] \tag{26}$$

$$F(R_m) = \frac{\mathrm{SSR}_m/(r-1)}{\mathrm{SSE}/(n-r)} \sim F[r-1, n-r, \boldsymbol{\beta}^{*\prime}\mathbf{X}_1'(\mathbf{I}-n^{-1}\mathbf{1}\mathbf{1}')\mathbf{X}_1\boldsymbol{\beta}^*/2\sigma^2] \tag{27}$$

$$F(R_m) = \frac{\mathrm{SSR}_m/(r-1)}{\mathrm{SSE}/(n-r)} \sim F(r-1, n-r, \boldsymbol{\beta}^{*\prime}\mathscr{A}'\mathscr{A}\,\boldsymbol{\beta}^*/2\sigma^2) \tag{28}$$

Cu acestea sîntem în măsură să întocmim

Tabelul 1

Analiza dispersiei pentru ajustarea modelului

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Sursa variației	SS	Gr. de libertate	$\overline{\mathrm{SS}}$	Statistica F
Modelul	$\mathrm{SSR} = \mathbf{b}^{0\prime}\mathbf{X}'\mathbf{Y}$	$n = r(\mathbf{X})$	$\overline{\mathrm{SSR}} = \frac{\mathrm{SSR}}{r}$	$F(R) = \frac{\overline{\mathrm{SSR}}}{\overline{\mathrm{SSE}}}$
Eroarea reziduală	$\mathrm{SSE} = \mathbf{Y}'\mathbf{Y} - \mathbf{b}^{0\prime}\mathbf{X}'\mathbf{Y}$	$n-r$	$\overline{\mathrm{SSE}} = \frac{\mathrm{SSE}}{n-r}$	
Total	$\mathrm{SST} = \mathbf{Y}'\mathbf{Y}$	n		

Scanned with OKEN Scanner

Tabelul 2

Analiza dispersiei pentru ajustarea modelului

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

a. Forma completă

Sursa variației	SS	Gr. de libertate	$\overline{SS}$	Statisticile F
Media	$SSM = n\bar{Y}^2$	1	$\overline{SSM} = \frac{SSM}{1}$	$F(M) = \frac{\overline{SSM}}{\overline{SSE}}$
Modelul (după ajustarea mediei)	$SSR_m = \mathbf{b}^{0\prime}\mathbf{X}'\mathbf{Y} - n\bar{Y}^2$	$r - 1$	$\overline{SSR}_m = \frac{SSR_m}{r-1}$	$F(R_m) = \frac{\overline{SSR}_m}{\overline{SSE}}$
Eroarea reziduală	$SSE = \mathbf{Y}'\mathbf{Y} - \mathbf{b}^{0\prime}\mathbf{X}'\mathbf{Y}$	$n - r$	$\overline{SSE} = \frac{SSE}{n-r}$	
Total	$SST = \mathbf{Y}'\mathbf{Y}$	n		

b. Forma prescurtată

Modelul (după ajustarea mediei)	$SSR_m = \mathbf{b}^{0\prime}\mathbf{X}'\mathbf{Y} - n\bar{Y}^2$	$r - 1$	$\overline{SSR}_m = \frac{SSE_m}{r-1}$	$F(R_m) = \frac{\overline{SSR}_m}{\overline{SSE}}$
Eroarea reziduală	$SSE = \mathbf{Y}'\mathbf{Y} - \mathbf{b}^{0\prime}\mathbf{X}'\mathbf{Y}$	$n - r$	$\overline{SSE} = \frac{SSE}{n-r}$	
Total (după ajustarea mediei)	$SST_m = \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2$	$r - 1$		

Observație. Tabelul 2*b* arată forma prescurtată a analizei dispersiei din tabelul 2*a*. Am obținut această formă prescurtată scăzînd din SST, SSM în scopul obținerii statisticii SST_m (ca în (14)). Astfel tabelul 2 *b*. nu conține $F(M)$, însă este identic cu tabelul 2 *a* în ceea ce privește $F(R_m) = \overline{SSR}_m/\overline{SSE}$.

2.5. Verificarea ipotezelor

Parametrul de necentricitate $\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}/2\sigma^2$ al repartiției F din (25) este zero în ipoteza nulă $\mathbf{X\beta} = \mathbf{0}$. Atunci $F(R) \sim F(r, n-r)$ și poate fi folosită ca bază a unui test pentru verificarea ipotezei nule. Cînd $F(R)$ este semnificativ putem spune că există concordanță între date și modelul $M(\mathbf{Y}) = \mathbf{X\beta}$. Aceasta nu arată că modelul folosit este cel mai convenabil; se poate folosi o submulțime de elemente care este la fel de semnificativă

Scanned with OKEN Scanner

ca întreaga mulțime : sau pot fi modele neliniare care sînt cel puțin la fel de convenabile ca modelul folosit.

Observăm că în contrast cu cazul modelului de rang maxim, testul bazat pe $F(R)$ nu poate fi folosit pentru verificarea ipotezei $H : \boldsymbol{\beta} = \mathbf{0}$, deoarece, după cum vom arăta, $\boldsymbol{\beta}$ nu este o *funcție estimabilă* și aceasta arată că $H : \boldsymbol{\beta} = \mathbf{0}$ nu poate fi verificată. Totuși $H : \mathbf{X}\boldsymbol{\beta} = \mathbf{0}$ poate fi verificată și $F(R)$ este o statistică corespunzătoare.

Parametrul de necentricitate $(\mathbf{1}'\mathbf{X}\boldsymbol{\beta})^2/2n\sigma^2$ al repartiției F din (26) este egal cu $n[M(\bar{Y})]^2/2\sigma^2$. Acest parametru este egal cu zero dacă este adevărată ipoteza $H : M(\bar{Y}) = 0$ caz în care $F(M) \sim F(1, n - r)$ și poate fi folosită ca bază a unui test pentru verificarea ipotezei $H : M(\bar{Y}) = 0$. Testul este bazat pe compararea statisticii $F(M)$ cu valorile tabelate ai variabilei $F(1, n - r)$. Un test echivalent se obține, comparînd $\sqrt{F(M)}$ cu valorile tabelate ale variabilei $t(n - r)$. Verificarea ipotezei $M(\bar{Y}) = 0$ poartă numele de verificarea mediei.

Pentru același motiv pentru care $F(R)$ nu poate fi o statistică test pentru $H : \boldsymbol{\beta} = \mathbf{0}$, nici $F(R_m)$ nu poate fi o statistică test pentru $H : \boldsymbol{\beta}^* = \mathbf{0}$, deoarece după cum vom arăta, $\boldsymbol{\beta}^*$ nu este *o funcție estimabilă* și astfel $H : \boldsymbol{\beta}^* = \mathbf{0}$ nu poate fi verificată.

Prin urmare, $F(R_m)$ poate fi considerată ca o statistică test pentru modelul $M(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ față de $M(Y) = \beta_0$.

Cînd $F(R_m)$ este semnificativ, aceasta nu înseamnă că toate elementele lui $\boldsymbol{\beta}^*$ sînt nenule, ci numai că unul dintre ele, sau o combinație liniară de ele poate fi. Dacă mai întîi a fost găsit că $F(M)$ este semnificativ, atunci $F(R_m)$ fiind semnificativ indică că modelul cu termeni în plus explică semnificativ mai mult variația în variabila Y decît o face modelul $M(Y) = \beta_0$.

Testele ce folosesc $F(M)$ și $F(R_m)$ sînt bazate pe numărători care sînt independenți, cu toate că numitorii lor sînt identici. Prin urmare statisticile F nu sînt independente.

§ 3. Funcții estimabile

3.1. Definiția funcției estimabile

DEFINIȚIA 2. Vom spune că $\mathbf{q}'\boldsymbol{\beta}$ este estimabilă dacă

$$\mathbf{q}'\boldsymbol{\beta} = \mathbf{t}'M(\mathbf{Y}) \tag{29}$$

pentru un vector $\mathbf{t}$, sau cu alte cuvinte : dacă există un vector $\mathbf{t}'$ astfel încît

$$\mathbf{t}'M(\mathbf{Y}) = \mathbf{q}'\boldsymbol{\beta},$$

atunci vom spune că $\mathbf{q}'\boldsymbol{\beta}$ este estimabilă.

Observație. Valoarea lui $\mathbf{t}'$ nu este atît de importantă ca existența sa și în acest sens pentru a stabili faptul că $\mathbf{q}'\boldsymbol{\beta}$ este estimabilă trebuie să vedem dacă există cel puțin o funcție liniară de valorile medii ale lui $\mathbf{Y}$, $\mathbf{t}'M(\mathbf{Y})$, a cărei valoare este $\mathbf{q}'\boldsymbol{\beta}$. Deoarece

$$\mathbf{t}'M(\mathbf{Y}) = M(\mathbf{t}'\mathbf{Y})$$

Scanned with OKEN Scanner

aceasta este echivalent cu stabilirea unei funcţii liniare de $\mathbf{Y}$, a cărei valoare medie este $\mathbf{q}'\boldsymbol{\beta}$. În general există mai multe astfel de funcţii $\mathbf{Y}$, stabilirea existenţei uneia dintre ele fiind suficientă pentru stabilirea estimabilităţii.

3.2. Proprietăţile funcţiilor estimabile

TEOREMA 5. i) *Valoarea medie a oricărei observaţii este estimabilă.*

ii) *Orice combinaţie liniară de funcţii estimabile este estimabilă*

iii) $\mathbf{q}'\boldsymbol{\beta}$ *este estimabilă dacă şi numai dacă* $\mathbf{q}' = \mathbf{t}'\mathbf{X}$

iv) *Dacă* $\mathbf{q}'\boldsymbol{\beta}$ *este estimabilă, atunci* $\mathbf{q}'\mathbf{b}^0$ *este un invariant sub* $\mathbf{G}$ *şi deci sub* $\mathbf{b}^0$.

v) $\mathbf{q}'\mathbf{b}^0$ *este cel mai bun estimator liniar nedeplasat pentru* $\mathbf{q}'\boldsymbol{\beta}$.

$$\widehat{\mathbf{q}'\boldsymbol{\beta}} = \mathbf{q}'\mathbf{b}^0. \tag{30}$$

Demonstraţie. Considerăm un $\mathbf{t}'$ care are un element egal cu unitatea şi celelalte zero. $\mathbf{t}'M(\mathbf{Y})$ va fi estimabilă şi va fi un element al lui $M(\mathbf{Y})$, adică valoarea medie a unei observaţii.

i) este demonstrată.

ii) rezultă din faptul că orice (sau fiecare) funcţie estimabilă este o combinaţie liniară de elementele lui $M(\mathbf{Y})$. Prin urmare este combinaţie liniară de funcţii estimabile. Astfel, dacă $\mathbf{q}_1'\boldsymbol{\beta}$ şi $\mathbf{q}_2'\boldsymbol{\beta}$ sînt estimabile atunci $\mathbf{q}_1'\boldsymbol{\beta} = \mathbf{t}_1'M(\mathbf{Y})$ şi $\mathbf{q}_2'\boldsymbol{\beta} = \mathbf{t}'M_2(\mathbf{Y})$ pentru anumiţi $\mathbf{t}_1$ şi $\mathbf{t}_2$.
Deci

$$c_1\mathbf{q}_1'\boldsymbol{\beta} + c_2\mathbf{q}_2'\boldsymbol{\beta} = (c_1\mathbf{t}_1' + c_2\mathbf{t}_2')M(\mathbf{Y})$$

şi astfel este estimabilă.

Dacă $\mathbf{q}'\boldsymbol{\beta}$ este estimabilă, $\mathbf{q}'\boldsymbol{\beta} = \mathbf{t}'M(\mathbf{Y})$ (prin definiţie) şi astfel $\mathbf{q}'\boldsymbol{\beta} = \mathbf{t}'\mathbf{X}\boldsymbol{\beta}$. Deoarece estimabilitatea este un concept care nu depinde de valoarea lui $\boldsymbol{\beta}$, acest ultim rezultat trebuie să aibă loc pentru toate valorile lui $\boldsymbol{\beta}$. Prin urmare :

$$\mathbf{q}' = \mathbf{t}'\mathbf{X} \tag{31}$$

pentru unii $\mathbf{t}'$. Pentru orice funcţie estimabilă $\mathbf{q}'\boldsymbol{\beta}$ nu este importantă o valoare specificată pentru $\mathbf{t}'$ ci existenţa unui $\mathbf{t}'$ ce satisface (31). ii) este demonstrată.

Cînd $\mathbf{q}'\boldsymbol{\beta}$ este estimabilă, $\mathbf{q}'\mathbf{b}^0$ este aceeaşi indiferent de soluţia ecuaţiei $\mathbf{X}'\mathbf{X}\mathbf{b}^0 = \mathbf{X}'\mathbf{Y}$ pentru $\mathbf{b}^0$, deoarece, după (31)

$$\mathbf{q}'\mathbf{b}^0 = \mathbf{t}'\mathbf{X}\mathbf{b}^0 = \mathbf{t}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{Y}$$

şi $\mathbf{X}\mathbf{G}\mathbf{X}'$ este invariantă faţă de $\mathbf{G}$ (problema 6). Prin urmare $\mathbf{q}'\mathbf{b}^0$ este invariantă faţă de $\mathbf{G}$ şi deci faţă de $\mathbf{b}^0$ (cînd $\mathbf{q}'\boldsymbol{\beta}$ este estimabilă). iv) este demonstrată.

Scanned with OKEN Scanner

Deoarece $\mathbf{q}'\mathbf{b}^0 = \mathbf{q}'\mathbf{G}\mathbf{X}'\mathbf{Y}$, $\mathbf{q}'\mathbf{b}^0$ este o funcţie liniară de observaţii Avem (folosim (31))

$$M(\mathbf{q}'\mathbf{b}^0) = \mathbf{q}'M(\mathbf{b}^0) = \mathbf{q}'\mathbf{H}\boldsymbol{\beta} = \mathbf{t}'\mathbf{X}\mathbf{H}\boldsymbol{\beta} = \mathbf{q}'\boldsymbol{\beta}$$

ceea ce arată că $\mathbf{q}'\mathbf{b}^0$ este un estimator nedeplasat pentru $\mathbf{q}'\boldsymbol{\beta}$.

Din problema 6

(32) $$\mathbf{X} = \mathbf{X}\mathbf{H} = \mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X} \text{ care implică } \mathbf{X}' = \mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{X}'.$$

Pentru a demonstra că $\mathbf{q}'\mathbf{b}^0$ este cel mai bun estimator avem nevoie de dispersia sa. Ţinînd seama de (5), (31), (32) şi (31) găsim

(33)
$$\begin{aligned} D^2(\mathbf{q}'\mathbf{b}^0) &= \mathbf{q}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{q}\,\sigma^2 \\ &= \mathbf{q}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{X}'\mathbf{t}\,\sigma^2 \\ &= \mathbf{q}'\mathbf{G}\mathbf{X}'\mathbf{t}\,\sigma^2 \\ &= \mathbf{q}'\mathbf{G}\mathbf{q}\,\sigma^2. \end{aligned}$$

Arătăm că printre toţi estimatorii liniari nedeplasaţi, $\mathbf{q}'\mathbf{b}^0$ are dispersia minimă şi deci că este cel mai bun. Presupunem că $\mathbf{a}'\mathbf{Y}$ este un alt estimator liniar nedeplasat pentru $\mathbf{q}'\boldsymbol{\beta}$ diferit de $\mathbf{q}'\mathbf{b}^0$. Atunci, deoarece $\mathbf{a}'\mathbf{Y}$ este nedeplasat

$$M(\mathbf{a}'\mathbf{Y}) = \mathbf{q}'\boldsymbol{\beta}$$

şi astfel

$$\mathbf{a}'\mathbf{X} = \mathbf{q}'.$$

Prin urmare

$$\operatorname{cov}(\mathbf{q}'\mathbf{b}^0, \mathbf{a}'\mathbf{Y}) = \operatorname{cov}(\mathbf{q}'\mathbf{G}\mathbf{X}'\,\mathbf{Y}, \mathbf{a}'\mathbf{Y}) = \mathbf{q}'\mathbf{G}\mathbf{X}'\mathbf{a}\,\sigma^2 = \mathbf{q}'\mathbf{G}\mathbf{q}\,\sigma^2.$$

În consecinţă

(34)
$$\begin{aligned} D^2(\mathbf{q}'\mathbf{b}^0 - \mathbf{a}'\mathbf{Y}) &= D^2(\mathbf{q}'\mathbf{b}^0) + D^2(\mathbf{a}'\mathbf{Y}) - 2\operatorname{cov}(\mathbf{q}'\mathbf{b}^0, \mathbf{a}'\mathbf{Y}) = \\ &= D^2(\mathbf{a}'\mathbf{Y}) - \mathbf{q}'\mathbf{G}\mathbf{q}\,\sigma^2 = \\ &= D^2(\mathbf{a}'\mathbf{Y}) - D^2(\mathbf{q}'\mathbf{b}^0). \end{aligned}$$

Însă $D^2(\mathbf{q}'\mathbf{b}^0 - \mathbf{a}'\mathbf{Y})$ este pozitivă şi prin urmare, din (34), $D^2(\mathbf{a}'\mathbf{Y})$ depăşeşte $D^2(\mathbf{q}'\mathbf{b}^0)$; cu alte cuvinte $\mathbf{q}'\mathbf{b}^0$ are o dispersie mai mică decît a oricărui estimator liniar nedeplasat pentru $\mathbf{q}'\boldsymbol{\beta}$ şi astfel este cel mai bun. Teorema este demonstrată.

Observaţie. Covarianţa dintre cele mai bune estimaţii liniare nedeplasate pentru două funcţii estimabile este

(35) $$\operatorname{cov}(\mathbf{q}_1'\mathbf{b}^0, \mathbf{q}_2'\mathbf{b}^0) = \mathbf{q}_1'\mathbf{G}\mathbf{q}_2\,\sigma^2$$

şi deci dacă $\mathbf{Q}'\mathbf{b}^0$ reprezintă cel mai bun estimator liniar nedeplasat pentru mai multe funcţii estimabile, atunci

(36) $$\mathscr{C}(\mathbf{Q}'\mathbf{b}^0) = \mathbf{Q}'\mathbf{G}\mathbf{Q}\,\sigma^2$$

Scanned with OKEN Scanner

3.3. Intervale de încredere

Deoarece numai funcţiile estimabile au estimatori liniari nedeplasaţi de dispersie minimă care sînt invarianţi la soluţia ecuaţiilor normale, ele sînt singurele funcţii pentru care este posibilă construirea intervalelor de încredere. La fel cu (XXXI.64) obţinem

COROLARUL 2. *Intervalul*

$$(37) \qquad (\mathbf{q}'\mathbf{b}^0 - \hat{\sigma} t_{1-\frac{\alpha}{2}}(n-r)\sqrt{\mathbf{q}'\mathbf{G}\mathbf{q}},\ \mathbf{q}'\mathbf{b}^0 + \hat{\sigma} t_{1-\frac{\alpha}{2}}(n-r)\sqrt{\mathbf{q}'\mathbf{G}\mathbf{q}})$$

este un interval de încredere pentru funcţia estimabilă $\mathbf{q}'\boldsymbol{\beta}$ *avînd coeficientul de încredere* $1-\alpha$.

Demonstraţia rezultă din teoremele 3 şi 5 şi din § 2 Cap. XII.

Observaţie. Dacă $N-r$ are valori mari ($\geqslant 100$), intervalul (37) devine

$$(37') \qquad (\mathbf{q}'\mathbf{b}^0 - \hat{\sigma} z_{1-\frac{\alpha}{2}}\sqrt{\mathbf{q}'\mathbf{G}\mathbf{b}},\ \mathbf{q}'\mathbf{b}^0 + \hat{\sigma} z_{1-\frac{\alpha}{2}}\sqrt{\mathbf{q}'\mathbf{G}\mathbf{q}}).$$

Ori de cîte ori $\mathbf{q}' = \mathbf{t}'\mathbf{X}$ pentru anumiţi $\mathbf{t}$, $\mathbf{q}'\boldsymbol{\beta}$ este estimabilă, cu cel mai bun estimator liniar avînd dispersia $\mathbf{q}'\mathbf{G}\mathbf{q}\,\sigma^2$.

3.4. Funcţii ce sînt estimabile

PROPOZIŢIA 3. a) *Orice funcţie* $\mathbf{m}'\mathbf{X}\boldsymbol{\beta}$ *pentru orice vector* $\mathbf{m}'$ *cu* n *componente este estimabilă cu*

$$(38) \qquad \widehat{\mathbf{m}'\mathbf{X}\boldsymbol{\beta}} = \mathbf{m}'\mathbf{X}\mathbf{b}^0 = \mathbf{m}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{Y}$$

şi

$$(39) \qquad D^2(\widehat{\mathbf{m}'\mathbf{X}\boldsymbol{\beta}}) = \mathbf{m}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{m}\,\sigma^2$$

b) Orice funcţie liniară de $\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$, $\mathbf{s}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$ (unde $\mathbf{s}'$ este orice vector de ordin p) este estimabilă cu

$$(40) \qquad \widehat{\mathbf{s}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}} = \mathbf{s}'\mathbf{X}'\mathbf{Y}$$

şi

$$(41) \qquad D^2(\widehat{\mathbf{s}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}) = \mathbf{s}'\mathbf{X}'\mathbf{X}\mathbf{s}\,\sigma^2$$

Observaţii. i) Deoarece $\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$ este acelaşi cu membrul stîng al ecuaţiilor normale cu $\boldsymbol{\beta}$ înlocuit cu $\mathbf{b}^0$ şi $\mathbf{s}'\mathbf{X}'\mathbf{Y}$, unde $\mathbf{X}'\mathbf{Y}$ este membrul drept al ecuaţiilor normale, este cel mai bun estimator liniar nedeplasat pentru

Scanned with OKEN Scanner

$\mathbf{s'X'X\beta}$, putem spune că cel mai bun estimator liniar nedeplasat pentru orice funcţie liniară de membrii stîngi ai ecuaţiilor normale este aceeaşi funcţie de membrii din dreapta.

ii) Dacă definim $\mathbf{x}_j$ ca coloana j a lui $\mathbf{X}$, atunci

$$\mathbf{X} \equiv (\mathbf{x}_1\, \mathbf{x}_2 \ldots \mathbf{x}_p)$$

şi

$$\mathbf{m'X\beta} = (\mathbf{m'x}_1)\beta_1 + (\mathbf{m'x}_2)\beta_2 + \ldots + (\mathbf{m'x}_p)\beta_p \tag{42}$$

cu

$$\mathbf{m'X\beta} = \mathbf{m'Xb}^0 = (\mathbf{m'x}_1)\mathbf{b}_1^0 + (\mathbf{m'x}_2)\mathbf{b}_2^0 + \ldots + (\mathbf{m'x}_p)b_p^0. \tag{43}$$

Pentru orice valori date m_t ale elementelor lui $\mathbf{m}$ folosite în (42) se obţine o funcţie estimabilă şi pentru valori folosite în (43) se obţine cel mai bun estimator liniar nedeplasat al funcţiei estimabile. Deci (42) şi (43) constituie expresii generale pentru o funcţie estimabilă şi respectiv pentru cel mai bun estimator liniar nedeplasat al ei.

iii) Dacă definim $\mathbf{z}_j$ ca coloana de ordin j a lui $\mathbf{X'X}$, atunci

$$\mathbf{X'X} = (\mathbf{Z}_1\mathbf{Z}_2 \ldots \mathbf{Z}_p)$$

şi

$$\mathbf{s'X'X\beta} = (\mathbf{s'Z}_1)\beta_1 + (\mathbf{s'Z}_2)\beta_2 + \ldots + (\mathbf{s'Z}_p)\beta_p \tag{44}$$

cu

$$\widehat{\mathbf{s'X'X\beta}} = \mathbf{s'X'Xb}^0 = (\mathbf{s'Z}_1)\mathbf{b}_1^0 + (\mathbf{s'Z}_2)b_2^0 + \ldots + (\mathbf{s'Z}_p)b_p^0. \tag{45}$$

Aceste expresii au loc pentru orice elemente ale lui $\mathbf{s'}$ de ordin p.

PROPOZIŢIA 4. a) *Orice funcţie liniară de* $M(\mathbf{b}^0)$, $\mathbf{u'}M(\mathbf{b}^0)$ ($\mathbf{u'}$ *arbitrar*) *este estimabilă cu*

$$\widehat{\mathbf{u'}M(\mathbf{b}^0)} = \mathbf{u'GX'Y} = \mathbf{u'b}^0 \tag{46}$$

şi

$$D^2[\widehat{\mathbf{u'}M(\mathbf{b}^0)}] = D^2(\mathbf{u'b}^0) = \mathbf{u'GX'XG'u}\,\sigma^2 \tag{47}$$

b) $\widehat{\mathbf{H\beta}} = \mathbf{b}^0$ şi

$$D^2(\mathbf{b}^0) = \mathbf{GX'XG'}\,\sigma^2 \tag{48}$$

Observaţie. Din punctul b) urmează că $\mathbf{w'H\beta}$ este de asemenea estimabilă cu cel mai bun estimator liniar nedeplasat $\mathbf{w'b}^0$. Astfel dacă

$$\mathbf{w'} = (w_1\, w_2 \ldots w_p)$$

şi

$$\mathbf{H} = (\mathbf{h}_1\mathbf{h}_2 \ldots \mathbf{h}_p)$$

Scanned with OKEN Scanner

atunci o funcţie estimabilă este

$$\mathbf{w'H\beta} = (\mathbf{w'h}_1)\beta_1 + (\mathbf{w'h}_2)\beta_2 + \ldots + (\mathbf{w'h}_p)\beta_p \tag{49}$$

şi

$$\widehat{\mathbf{w'H\beta}} = \mathbf{w'b}^0 = w_1 b_1^0 + w_2 b_2^0 + \ldots + w_p b_p^0. \tag{50}$$

Expresiile (49) şi (50) au avantaj asupra expresiilor :

— (42) şi (43) bazate pe $\mathbf{m'X\beta}$ deoarece au mai puţine elemente arbitrare ; p în loc de n,

— (44) şi (45) din cauză că sînt mai simple. Acest fapt este evident în (50), care este o combinaţie liniară de elementele lui $\mathbf{b}^0$, fiecare inmulţit cu un singur element arbitrar w. Şi (49) are o formă simplă, cînd $\mathbf{X'X}$ este o matrice „plan". $\mathbf{H}$ are adeseori $p-r$ linii nule, $r = r(\mathbf{X})$ cu celelalte r linii avînd ca elemente 0, 1 sau -1. În consecinţă funcţia estimabilă din (49) ia o formă simplă şi conţine numai r elemente ale lui $\mathbf{w}$. În astfel de cazuri $\mathbf{b}^0$ poate avea numai r elemente nenule şi astfel în (50) avem numai r termeni.

Că $\mathbf{H}$ poate fi obţinută ca o matrice de 0, 1 şi -1 cînd $\mathbf{X'X}$ este o matrice plan se stabileşte după cum urmează. Presupunem că

$$\mathbf{X'X} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix} \text{ şi } \mathbf{G} = \begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{51}$$

unde $\mathbf{X}_1'\mathbf{X}_1$ are rang maxim, egal cu $r(\mathbf{X})$ şi $\mathbf{G}$ este inversa generalizată a lui $\mathbf{X'X}$. Deoarece $\mathbf{X} = (\mathbf{X}_1\ \mathbf{X}_2)$ unde $\mathbf{X}_1$ are rang maxim coloană $\mathbf{X}_2 = \mathbf{X}_1\mathbf{M}$ deoarece toate elementele lui $\mathbf{X}$ sînt 0 sau 1, acelea ale lui $\mathbf{M}$ pot fi 0, 1 sau -1. Deci

$$\mathbf{H} = \mathbf{GX'X} = \begin{pmatrix} \mathbf{I} & (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{M} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{52}$$

şi astfel $p-r$ linii ale lui $\mathbf{H}$ sînt nule şi elementele în celelalte r linii nenule sînt 0, 1 sau -1.

Observaţie. Din propoziţiile anterioare rezultă că există o infinitate de funcţii estimabile.

PROPOZIŢIA 5. *Există $r(\mathbf{X})$ funcţii estimabile liniar independente.*

Demonstraţie. Deoarece $\mathbf{q'\beta}$ cu $\mathbf{q}' = \mathbf{t'X}$ este estimabilă, fie $\mathbf{T}'$ o matrice $n \times n$ de rang maxim. Atunci, cu $\mathbf{Q}' = \mathbf{T'X}$, funcţiile $\mathbf{Q'\beta}$ sînt n funcţii estimabile. Însă $r(\mathbf{Q}) = r(\mathbf{X})$. Prin urmare în $\mathbf{Q}'$ sînt numai $r(\mathbf{X})$ linii liniar independente şi deci în $\mathbf{Q'\beta}$ sînt numai $r(\mathbf{X})$ termeni liniari independenţi ; adică numai $r(\mathbf{X})$ funcţii estimabile liniar independente. Astfel orice mulţime de funcţii estimabile nu poate conţine mai mult de r astfel de funcţii liniar independente.

O funcţie $\mathbf{q'\beta}$ dată este estimabilă dacă poate fi găsit un vector $\mathbf{t}'$ astfel încît $\mathbf{t'X} = \mathbf{q}'$. Totuşi, pentru $\mathbf{q}'$ cunoscut obţinerea lui $\mathbf{t}'$ care să satisfacă $\mathbf{t'X} = \mathbf{q}'$ nu este întotdeauna uşoară, mai ales cînd $\mathbf{X}$ este de dimensiuni mari. Alternativ obţinerii lui $\mathbf{t}'$, poate fi cercetată estimabilitatea lui $\mathbf{q'\beta}$. Ne interesează dacă $\mathbf{q}'$ este astfel încît ecuaţia $\mathbf{q'H} = \mathbf{q}'$ este satisfăcută.

Scanned with OKEN Scanner

PROPOZIȚIA 6. $\mathbf{q'\beta}$ *este estimabilă dacă și numai dacă* $\mathbf{q'H = q'}$.

Demonstrație. Dacă $\mathbf{q'\beta}$ este estimabilă, $\mathbf{q' = t'X}$ și $\mathbf{q'H = t'XH = t'X = q'}$.

Dacă $\mathbf{q'H = q'}$, atunci $\mathbf{q' = q'GX'X = t'X}$ pentru $\mathbf{t' = q'GX'}$.

Observație. Propoziția 6 dă un procedeu direct pentru a vedea dacă $\mathbf{q'\beta}$ este estimabilă; vedem dacă $\mathbf{q'H}$ este egal cu $\mathbf{q'}$. Cînd $\mathbf{q'H}$ este egal cu $\mathbf{q'}$, nu numai că $\mathbf{q'\beta}$ este estimabilă însă din propoziția 4, cel mai bun estimator liniar nedeplasat pentru

$$\mathbf{q'\beta = q'H\beta}$$

este $\mathbf{q'b^0}$. Acestea corespund la proprietatea de invarianță a lui $\mathbf{q'b^0}$ pentru $\mathbf{q'H = q'}$ (a se vedea problema 5).

§ 4. Ipoteze liniare generale

4.1. Ipoteze verificabile

DEFINIȚIA 3. Vom spune că o ipoteză este verificabilă dacă poate fi exprimată cu ajutorul funcțiilor estimabile.

Prin urmare ipoteza verificabilă $H: \mathbf{K'\beta = m}$ este una în care

$$\mathbf{K'\beta} \equiv \{\mathbf{k_i'\beta}\} \qquad \text{pentru } i = 1, 2, \ldots, s$$

astfel încît $\mathbf{k_i'\beta}$ este estimabilă pentru toți i. Deci $\mathbf{k_i' = t_i'X}$ pentru unele valori $\mathbf{t_i'}$ și astfel

$$\mathbf{K' = T'X} \tag{53}$$

pentru o anumită matrice $(\mathbf{T'})_{s\times n}$. Mai mult, orice ipoteză este considerată numai în termeni de componentele sale liniar independente și astfel $(\mathbf{K'})_{s\times p}$ este întotdeauna de rang maxim egal cu numărul de linii s.

Deoarece se consideră că $\mathbf{K'\beta}$ este o mulțime de funcții estimabile:

$$\widehat{\mathbf{K'\beta}} = \mathbf{K'b^0}. \tag{54}$$

TEOREMA 6. *Avem*

$$M(\mathbf{K'b^0}) = \mathbf{K'\beta} \tag{55}$$

și

$$\mathscr{C}(\widehat{\mathbf{K'\beta}}) = \mathbf{K'GK}\sigma^2, \tag{56}$$

matricea $\mathbf{K'GK}$ *fiind nesingulară.*

Demonstrație. Din teorema 1

$$M(\mathbf{K'b^0}) = \mathbf{K'}M(\mathbf{b^0}) = \mathbf{K'H}$$

și din propoziția 6, obținem (55).

Scanned with OKEN Scanner

Din teorema 1 și din (54), găsim

$$\mathscr{C}(\widehat{\mathbf{K}'\boldsymbol{\beta}}) = \mathbf{K}'\mathscr{C}(\mathbf{b}^0)\mathbf{K}$$
$$= \mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{K}\sigma^2$$

sau ținînd seama de (53)

$$\mathscr{C}(\widehat{\mathbf{K}'\boldsymbol{\beta}}) = \mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{X}\mathbf{T}\sigma^2.$$

Folosind problema 6 și (53) găsim (56).

Rămîne să arătăm că $\mathbf{K}'\mathbf{G}\mathbf{K}$ este nesingulară. Deoarece funcțiile $\mathbf{K}'\boldsymbol{\beta}$ sînt estimabile $\mathbf{K}'$ poate fi reprezentată nu numai ca $\mathbf{T}'\mathbf{X}$ pentru anumite matrici $\mathbf{T}'$ ci și ca $\mathbf{S}'\mathbf{X}'\mathbf{X}$ pentru anumite matrici $\mathbf{S}$ de rang maxim s. Atunci cu

$$\mathbf{K}' = \mathbf{S}'\mathbf{X}'\mathbf{X} \text{ de ordin } s \times p \text{ și } r(\mathbf{K}') = s$$

unde $s \leqslant r$, $\mathbf{S}'$ și $\mathbf{S}'\mathbf{X}'$ au rang maxim s (numărul de linii).

De exemplu, $r(\mathbf{S}') \geqslant r(\mathbf{K}')$ și astfel $r(\mathbf{S}') \geqslant s$; însă $\mathbf{S}'$ este $s \times p$ și astfel $r(\mathbf{S}') \leqslant s$; deci $r(\mathbf{S}') = s$. Mai mult,

$$\mathbf{K}'\mathbf{G}\mathbf{K} = \mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{S} = \mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S}$$

și astfel $r(\mathbf{K}'\mathbf{G}\mathbf{K}) = r(\mathbf{S}'\mathbf{X}') = s$ ordinul lui $\mathbf{K}'\mathbf{G}\mathbf{K}$. Deci $\mathbf{K}'\mathbf{G}\mathbf{K}$ nu este singulară. Teorema este demonstrată.

Verificăm ipoteza

$$H : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m} \tag{57}$$

ca în cazul rangului maxim.

Au loc

TEOREMA 7. *In ipoteza normalității vectorului* $\boldsymbol{\varepsilon}$,

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

$$\mathbf{b}^0 \sim \mathrm{N}(\mathbf{G}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}, \mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\sigma^2)$$

și

$$\mathbf{K}'\mathbf{b}^0 - \mathbf{m} \sim \mathrm{N}(\mathbf{K}'\boldsymbol{\beta} - \mathbf{m}, \mathbf{K}'\mathbf{G}\mathbf{K}\sigma^2).$$

TEOREMA 8. *Avem*

$$Q/\sigma^2 \sim \chi^2[s, (\mathbf{K}'\boldsymbol{\beta} - \mathbf{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\boldsymbol{\beta} - \mathbf{m})/2\sigma^2]$$

unde

$$Q = (\mathbf{K}'\mathbf{b}^0 - \mathbf{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})(\mathbf{K}'\mathbf{b}^0 - \mathbf{m}). \tag{58}$$

Demonstrația rezultă din problema 11.

Mai mult.

$$Q = [\mathbf{Y} - \mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m}]'\mathbf{X}\mathbf{G}'\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}\mathbf{K}'\mathbf{G}\mathbf{X}'[\mathbf{Y} - \mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m}],$$

Scanned with OKEN Scanner

cu $(\mathbf{K}'\mathbf{K})^{-1}$ existînd deoarece $\mathbf{K}'$ este de rang maxim s, și

$$\mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m} = \mathbf{T}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m} =$$

$$= \mathbf{T}'\mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m} = \mathbf{K}'\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m} = \mathbf{m}.$$

Deoarece $\mathbf{X}'(\mathbf{I} - \mathbf{X}\mathbf{G}\mathbf{X}') = \mathbf{0}$,

$$\text{SSE} = [\mathbf{Y} - \mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m}]'(\mathbf{I} - \mathbf{X}\mathbf{G}\mathbf{X}')[\mathbf{Y} - \mathbf{X}\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{m}]$$

și matricile formelor pătratice în aceste expresii au produsul nul. Astfel Q și SSE sînt independente. Prin urmare am demonstrat.

TEOREMA 9. *Avem*

$$F(H) = \frac{Q/s}{\text{SSE}/(n-r)} \sim F[s,\ n-r,\ (\mathbf{K}'\boldsymbol{\beta} - \mathbf{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\boldsymbol{\beta} - \mathbf{m})/2\sigma^2]$$

și în ipoteza nulă (57)

$$F(H) \sim F(s,\ n-r).$$

Cu ajutorul acestei teoreme, cu

$$F(H) = (\mathbf{K}'\mathbf{b}^0 - \mathbf{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^0 - \mathbf{m})/s\hat{\sigma}^2 \tag{59}$$

putem verifica ipoteza (57).

Ne propunem să determinăm, în cazul în care este adevărata ipoteză $H: \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$, valoarea pentru $\mathbf{b}^0$, pe care o notăm prin $\mathbf{b}^0_H$.

Vom minimiza

$$(\mathbf{Y} - \mathbf{X}\mathbf{b}^0_H)'(\mathbf{Y} - \mathbf{X}\mathbf{b}^0_H)$$

cu condiția $\mathbf{K}'\mathbf{b}^0_H = \mathbf{m}$.

La fel ca în (XXXI.74), folosind multiplicatorul lui Lagrange $2\boldsymbol{\theta}'$, obținem

$$\mathbf{X}'\mathbf{X}\mathbf{b}^0_H + \mathbf{K}\boldsymbol{\theta} = \mathbf{X}'\mathbf{Y}$$

$$\mathbf{K}'\mathbf{b}^0_H = \mathbf{m}. \tag{60}$$

Din prima ecuație a relației (60), găsim

$$\mathbf{b}^0_H = \mathbf{G}\mathbf{X}'\mathbf{Y} - \mathbf{G}\mathbf{K}\boldsymbol{\theta} = \mathbf{b}^0 - \mathbf{G}\mathbf{K}\boldsymbol{\theta}$$

de unde, prin înlocuire în a doua ecuație a relației (60) (ca în (XXXI. 74))

$$\mathbf{b}^0_H = \mathbf{b}^0 + \mathbf{G}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^0 - \mathbf{m}). \tag{61}$$

DEFINIȚIA 4. Statistica

$$\text{SSE}_H = (\mathbf{Y} - \mathbf{X}\mathbf{b}^0_H)'(\mathbf{Y} - \mathbf{X}\mathbf{b}^0_H) \tag{62}$$

poartă numele de SS erorilor după ajustare

Scanned with OKEN Scanner

Are loc

TEOREMA 10. *Avem*

$$\mathrm{SES}_H = \mathrm{SSE} + Q \tag{63}$$

unde Q *este dat de* (58).

Demonstrație. Putem să scriem

$$\mathrm{SSE}_H = [\mathbf{Y} - \mathbf{X}\mathbf{b}^0 + \mathbf{X}(\mathbf{b}^0 - \mathbf{b}_H^0)]'[\mathbf{Y} - \mathbf{X}\mathbf{b}^0 + \mathbf{X}(\mathbf{b}^0 - \mathbf{b}_H^0)] =$$

$$= (\mathbf{Y} - \mathbf{X}\mathbf{b}^0)'(\mathbf{Y} - \mathbf{X}\mathbf{b}^0) + (\mathbf{b}^0 - \mathbf{b}_H^0)'\mathbf{X}'\mathbf{X}(\mathbf{b}^0 - \mathbf{b}_H^0),$$

produsul mixt anulîndu-se deoarece

$$\mathbf{X}'(\mathbf{Y} - \mathbf{X}\mathbf{b}^0) = 0.$$

Dacă în (64) ținem seama de $\mathbf{b}^0 - \mathbf{b}_H^0$ (obținut din (61)), găsim

$$\mathrm{SSE}_H = \mathrm{SSE} + (\mathbf{K}'\mathbf{b}^0 - \mathbf{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}\mathbf{K}'\mathbf{G}'\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^0 - \mathbf{m}).$$

Avem $\mathbf{K}' = \mathbf{T}'\mathbf{X}$ și astfel

$$\mathbf{K}'\mathbf{G}'\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1} = \mathbf{T}'\mathbf{X}\mathbf{G}'\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1} = \mathbf{T}'\mathbf{X}\mathbf{G}\mathbf{X}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1} = \mathbf{I}$$

și

$$\mathbf{K}'\mathbf{G}'\mathbf{K} = \mathbf{T}'\mathbf{X}\mathbf{G}'\mathbf{X}'\mathbf{T} = \mathbf{T}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{T} = \mathbf{K}'\mathbf{G}\mathbf{K}.$$

Deci

$$\mathrm{SSE}_H = \mathrm{SSE} + (\mathbf{K}'\mathbf{b}^0 - \mathbf{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^0 - \mathbf{m}),$$

ceea ce demonstrează teorema.

Observație. Deoarece (64) este limitată numai în cazul în care $\mathbf{K}'\boldsymbol{\beta}$ este estimabilă și deoarece $\boldsymbol{\beta}$ și $\boldsymbol{\beta}_q$ nu sînt estimabile nu pot fi verificate ipotezele

$$H : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \quad \text{și} \quad H : \boldsymbol{\beta}_q = \mathbf{0}.$$

Testele descrise în XXXI § 4 bazate pe $F(R)$ și $F(R_m)$ nu pot fi folosite pentru a verifica ipoteze de această natură.

Un caz special de ipoteze generale

$$\mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$$

este cînd $\mathbf{m}$ este nul. În acest caz Q și $\mathbf{b}_H^0$ din (58) și (61) devin

$$Q = \mathbf{b}^{0\prime}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}\mathbf{K}'\mathbf{b}^0 \quad \text{și} \quad \mathbf{b}_H^0 = \mathbf{b}^0 - \mathbf{G}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}\mathbf{K}'\mathbf{b}^0 \tag{65}$$

cu

$Q = \mathrm{SSR}$ — *reducerea în* SS *datorită ajustării modelului redus*

Scanned with OKEN Scanner

Deci corespunzător tabelului XXXI.4 avem

Tabelul 3

Analiza dispersiei pentru verificarea ipotezei $\mathbf{K}'\boldsymbol{\beta}=\mathbf{0}$

Sursa variației	SS	Gr. de libertate	$\overline{SS}$	Statistica F
Modelul maxim	$SSR = \mathbf{b}^{0\prime}\mathbf{X}'\mathbf{Y}$	r	SSR/r	$\frac{SSR/r}{SSE/(n-r)}$ pt. modelul maxim
Ipoteza	$Q = \mathbf{b}^{0\prime}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}\mathbf{K}'\mathbf{b}^0$	s	Q/s	$\frac{Q/s}{SSE/(n-r)}$ pt. ipoteza $\mathbf{K}'\boldsymbol{\beta}=\mathbf{0}$
Modelul redus	$SSR - Q$	$r-s$	$\frac{SSR-Q}{r-s}$	$\frac{(SSR-Q)/(r-s)}{SSE/(n-r)}$ pt. modelul redus sub H
Eroarea reziduală	SSE	$n-r$	$\frac{SSE}{n-r}$	
Total	$SST = \mathbf{Y}'\mathbf{Y}$	n		

$r = r(\mathbf{X})$, $s = r(\mathbf{K}')$ cu $\mathbf{K}'$ avînd rangul egal cu numărul de linii. Rescriem tabelul 3 făcînd să intervină termenii : după ajustare.

Tabelul 4

Analiza dispersiei pentru verificarea ipotezei $\mathbf{K}'\boldsymbol{\beta}=\mathbf{0}$ după ajustarea mediei (d. a. m)

Sursa variației	SS	Gr. de libertate	Statistica F
Modelul maxim d.a.m.	$SSR_m = SSR - n\bar{Y}^2$	$r-1$	$\frac{SSR_m/(r-1)}{SSE/(n-r)}$ ptr. modelul maxim d.a.m.
Ipoteza	$Q = \mathbf{b}^{0\prime}\mathbf{K}(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}\mathbf{K}'\mathbf{b}$	s	$\frac{Q/s}{SSE/(n-r)}$ ptr. ipoteza $\mathbf{K}'\boldsymbol{\beta}=\mathbf{0}$ d. a. m.
Modelul redus d.a.m.	$SSR_m - Q$	$r-s-1$	$\frac{(SSR_m-Q)/(r-s-1)}{SSE/(n-r)}$ ptr. modelul redus d.a.m.
Eroarea reziduală	SSE	$n-r$	
Total d.a.m.	$SST_m = \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2$	$n-1$	

Observație. În modelele care nu sînt de rang maxim, $\mathbf{G}$ și $\mathbf{b}^0$ înlocuiesc pe $(\mathbf{X}'\mathbf{X})^{-1}$ și respectiv $\mathbf{b}$ din modelele de rang maxim. Deci modelul de rang maxim este un caz particular al modelului care nu este de rang

Scanned with OKEN Scanner

maxim. Astfel cînd $\mathbf{X'X}$ nu este singulară, $\mathbf{G} = (\mathbf{X'X})^{-1}$ şi $\mathbf{b} = \mathbf{b}^0$ şi toate rezultatele pentru modelul de rang maxim urmează din acelea pentru modelul care nu este de rang maxim.

4.2. Ipoteze care nu sînt verificabile

TEOREMA 11. *Dacă* $\mathbf{K'\beta}$ *nu este estimabilă, nu există un test pentru verificarea ipotezei* $H : \mathbf{K'\beta} = \mathbf{m}$.

Demonstraţie. Ecuaţiile rezultate în urma minimizării lui $(\mathbf{Y} - \mathbf{Xb}^0)'(\mathbf{Y} - \mathbf{Xb}^0)$ cu condiţia $\mathbf{K'b}^0 = \mathbf{m}$ sînt

$$\mathbf{X'Xb}_H^0 + \mathbf{K\theta} = \mathbf{X'Y} \quad \text{şi} \quad \mathbf{K'b}_H^0 = \mathbf{m} \tag{66}$$

unde $2\boldsymbol{\theta}'$ este multiplicatorul lui Lagrange. Considerăm ecuaţiile în $\mathbf{z}_1$:

$$\mathbf{K'(H - I)z}_1 = \mathbf{m} - \mathbf{K'GX'Y}. \tag{67}$$

Din problema 4, $\mathbf{(H - I)z}_1$ conţine $p-r$ elemente arbitrare. Acum deoarece $\mathbf{K'\beta}$ nu este estimabilă, $\mathbf{K'} \neq \mathbf{T'X}$ pentru orice $\mathbf{T'}$ şi astfel deoarece $\mathbf{X} = \mathbf{XGX'X}$ (problema 6),

$$\mathbf{K'} \neq \mathbf{(T'XG)X'X}$$

pentru orice $\mathbf{T'}$, adică liniile lui $\mathbf{K'}$ sînt liniar independente de acelea ale lui $\mathbf{X'X}$. Însă $\mathbf{X'X}$ are ordinul p şi rangul r. De asemenea, liniile lui $\mathbf{K'}$ au ordinul p şi sînt liniar indepenente unele de altele. Prin urmare, dacă ele sînt de asemenea liniar independente de liniile lui $\mathbf{X'X}$ nu pot fi mai mult de $p-r$; adică $\mathbf{K'}$ nu are mai mult de $p-r$ linii. Deci (67) nu are mai mult de $p-r$ ecuaţii în $p-r$ necunoscute ale lui $\mathbf{(H - I)z}_1$, şi astfel are cel puţin o soluţie pentru $\mathbf{z}_1$. Folosind-o pentru $\mathbf{z}$ din

$$\mathbf{b}^0 = \mathbf{GX'Y} + \mathbf{(H - I)z} \tag{68}$$

obţinem

$$\mathbf{b}_H^0 = \mathbf{GX'Y} + \mathbf{(H - I)z}_1, \tag{69}$$

şi găsim că $\boldsymbol{\theta} = \mathbf{0}$ şi $\mathbf{b}_H^0$ din (69) satisface (66).

În consecinţă, deoarece (69) este o submulţime a soluţiilor (68) ale lui $\mathbf{X'Xb}^0 = \mathbf{X'Y}$,

$$\mathrm{SSE}_H = (\mathbf{Y} - \mathbf{Xb}_H^0)'(\mathbf{Y} - \mathbf{Xb}_H^0) = \mathrm{SSE}$$

şi astfel nu există un test pentru verificarea ipotezei. Teorema este demonstrată.

Observaţii. i) Comparînd (60) cu (66) s ngura diferenţă dintre ele este că în (60) $\mathbf{K'\beta}$ este estimabilă în timp ce în (66) nu.

ii) Rezolvînd (60) avem condiţia de estimabilitate ($\mathbf{K'} = \mathbf{T'X}$ pentru anumiţi $\mathbf{T'}$) şi obţinem soluţia (61). Pe de altă parte, în rezolvarea lui (66), soluţia pentru $\mathbf{b}_H^0$ este de asemenea o soluţie, pentru $\mathbf{X'X\beta} = \mathbf{X'Y}$,

Scanned with OKEN Scanner

după cum am arătat în (67) și (69). Lipsa de estimabilitate a lui $\mathbf{K}'\boldsymbol{\beta}$ permite aceasta. În contrast, în (60) unde $\mathbf{K}'\boldsymbol{\beta}$ este estimabilă

$$\mathbf{K}' = \mathbf{S}'\mathbf{X}'\mathbf{X}$$

pentru anumiți $\mathbf{S}'$ și astfel, pentru $\mathbf{b}^0$ din (68)

$$\mathbf{K}'\mathbf{b}^0 = \mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{b}_H^0 = \mathbf{S}'\mathbf{X}\mathbf{Y}$$

pentru toate valorile lui $\mathbf{z}$. Prin urmare nu poate fi găsită nici o valoare a lui $\mathbf{z}$ din (68) astfel încît $\mathbf{K}'\mathbf{b}^0 = \mathbf{m}$ și astfel nu există nici o valoare a lui (68) care să satisfacă (60).

O extensie constă în a încerca să verificăm ipoteze care parțial constau din funcții estimabile și parțial din funcții care nu sînt estimabile. Presupunem că $H : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ poate fi scrisă ca

$$H : \begin{pmatrix} \mathbf{K}_1'\boldsymbol{\beta} \\ \mathbf{k}'\boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{pmatrix} \tag{70}$$

unde $\mathbf{K}_1'\boldsymbol{\beta}$ este estimabilă însă $\mathbf{k}'\boldsymbol{\beta}$ nu. Atunci, folosind doi multilpicatori ai lui Lagrange și procedînd ca mai înainte, ajungem la concluzia că verificarea ipotezei (70) nu se distinge de aceea a ipotezei $H : \mathbf{K}_1'\boldsymbol{\beta} = \mathbf{m}_1$.

Statistica

$$Q = (\mathbf{K}'\mathbf{b}^0 - \mathrm{m})'(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{b}^0 - \mathbf{m})$$

este legată de estimabilitatea lui $\mathbf{K}'\boldsymbol{\beta}$.

Cînd $\mathbf{K}'\boldsymbol{\beta}$ nu este estimabilă, Q poate fi calculată atîta timp cît $\mathbf{K}'\mathbf{G}\mathbf{K}$ nu este singulară. Este așa deoarece estimabilitatea lui $\mathbf{K}'\boldsymbol{\beta}$ este o condiție suficientă pentru existența lui Q în particular pentru existența lui $(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}$ însă ea nu este o condiție necesară. Deci, cînd $(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1}$ există, Q poate fi calculată chiar dacă $\mathbf{K}'\boldsymbol{\beta}$ nu este estimabilă. Prin urmare este important ca înainte de a calcula Q și $F(H)$ să vedem dacă $\mathbf{K}'\boldsymbol{\beta}$ este estimabilă. Pentru aceasta stabilim existența unui $\mathbf{T}'$ astfel încît $\mathbf{K} = \mathbf{T}'\mathbf{X}$ sau cercetăm dacă $\mathbf{K}'$ satisface $\mathbf{K}' = \mathbf{K}'\mathbf{H}$.

Totuși presupunem că stabilirea în acest mod a estimabilității lui $\mathbf{K}'\boldsymbol{\beta}$ este trecută cu vederea și este calculată Q. Atunci, dacă $\mathbf{K}'\boldsymbol{\beta}$ nu este estimabilă, ce ipoteză se verifică cu ajutorul statisticii $F(H)$? Are loc :

TEOREMA 12. *Dacă* $\mathbf{K}'\boldsymbol{\beta}$ *nu este estimabilă,* $F(H)$ *este bază a unui test pentru verificarea ipotezei* $H : \mathbf{K}'\mathbf{H}\boldsymbol{\beta} = \mathbf{m}$.

Demonstrație. Deoarece $H : \mathbf{K}'\mathbf{H}\boldsymbol{\beta} = \mathbf{m}$ este întotdeauna verificabilă, valoarea lui Q pentru verificarea acestei ipoteze (s-o notăm prin Q_1) este (din (58)).

$$Q_1 = (\mathbf{K}'\mathbf{H}\mathbf{b}^0 - \mathbf{m})'(\mathbf{K}'\mathbf{H}\mathbf{G}\mathbf{H}'\mathbf{K})^{-1}(\mathbf{K}'\mathbf{H}\mathbf{b}^0 - \mathbf{m}). \tag{71}$$

În această expresie

$$\mathbf{K}'\mathbf{H}\mathbf{b}^0 = \mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{Y} = \mathbf{K}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{X}'\mathbf{Y} = \mathbf{K}'\mathbf{G}_1\mathbf{X}'\mathbf{Y}$$

deoarece (după problema 6) $\mathbf{X}\mathbf{G}\mathbf{X}' = \mathbf{X}\mathbf{G}'\mathbf{X}'$ și unde

$$\mathbf{G}_1 = \mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'$$

Scanned with OKEN Scanner

este inversa generalizată a lui $\mathbf{X'X}$. Prin urmare

$$\mathbf{K'Hb^0 = K'G_1X'Y = K'b_1^0}$$

unde

$$\mathbf{b_1^0 = G_1X'Y}$$

este soluția ecuației $\mathbf{X'Xb^0 = X'Y}$. De asemenea,

$$\mathbf{K'HGH'K = K'GX'XGX'XG'K = K'GX'XG'K = K'G_1K}$$

și astfel din (71)

$$Q_1 = (\mathbf{K'b_1^0 - m})'(\mathbf{K'G_1K})^{-1}(\mathbf{K'b_1^0 - m}).$$

Astfel Q_1 este identic cu numărătorul SS calculat din (58) pentru verificarea ipotezei neverificabile $\mathbf{K'\beta = m}$ folosind soluța $\mathbf{b_1^0 = G_1X'Y}$. Deci calculele ce trebuiesc făcute cînd încercăm să verificăm ipoteza neverificabilă $\mathbf{K'\beta = m}$ nu se disting de acelea întîlnite în verificarea ipotezei testabile $\mathbf{K'H\beta = m}$; adică dacă $F(H)$ din (59) este calculată pentru ipoteza $\mathbf{K'\beta = m}$ care nu este verificabilă, ipoteza actuală care este verificată este $\mathbf{K'H\beta = m}$. Teorema este demonstrată.

Numărătorul SS pentru verificarea ipotezei $\mathbf{K'\beta = 0}$ este ca în (65)

$$Q = \mathbf{b^{0\prime}K(K'GK)^{-1}K'b^0}.$$

$\mathbf{K'\beta}$ fiind estimabilă, din teorema 6, $\mathbf{K' = S'X'X}$ pentru unii $\mathbf{S'}$ și astfel cu $\mathbf{b^0 = GX'Y}$,

$$Q = \mathbf{Y'XG'X'XS(S'X'XGX'XS)^{-1}S'X'XGX'Y = Y'XS(S'X'XS)^{-1}S'X'Y}$$

(folosind problema 6). Mai mult, $\mathbf{K'}$ are rang maxim linie : s, și cînd $s = r = r(\mathbf{X})$, $\mathbf{XS = X_1PX_1'X}$ unde $\mathbf{X_1}$ este o submatrice $n \times r$ a lui X de rang maxim coloană, cu $\mathbf{P}$ și $\mathbf{X_1'X}$ nesingulare. Aceasta arată că $\mathbf{S(S'X'XS)^{-1}S'}$, este inversa generalizată a lui $\mathbf{X'X}$ și astfel

(72) $$Q = \mathbf{Y'XGX'Y} = \text{SSR} \quad \text{cînd} \quad s = r = r(\mathbf{X}).$$

Prin urmare am demonstrat

TEOREMA 13. *În cazul în care* $\mathbf{K'\beta}$ *reprezintă numărul maxim de funcții estimabile liniar independente și anume* $r = r(\mathbf{X})$, SSR *datorită ajustării modelului* $M(\mathbf{Y}) = \mathbf{X\beta}$ *se reduce la numărătorul lui* SS *pentru verificarea ipotezei* $\mathbf{K'\beta = 0}$.

Observație. Teorema 13 arată că dacă $\mathbf{k_i'}$ este o linie a lui $\mathbf{K'}$, atunci numărătorul SS pentru verificarea simultană a ipotezelor $\mathbf{k_i'\beta} = 0$ pentru $i = 1, 2, \ldots, r$ este egal cu SSR. Însă aceasta nu înseamnă că pentru verificarea individuală a r ipoteze $\mathbf{k_i'\beta} = 0$, sumele pătratelor adunate dau SSR. Va fi așa numai în anumite cazuri.

Presupunem că $\mathbf{k_i'}$ și $\mathbf{k_j'}$ sînt două linii ale lui $\mathbf{K'}$.

Scanned with OKEN Scanner

Atunci

$$q_i = \mathbf{b}^{0\prime}\mathbf{k}_i(\mathbf{k}_i'\mathbf{G}\mathbf{k}_i)^{-1}\mathbf{k}_i'\mathbf{b}^0 = \mathbf{Y}'\mathbf{X}\mathbf{G}'\mathbf{k}_i(\mathbf{k}_i'\mathbf{G}\mathbf{k}_i)^{-1}\mathbf{k}_i'\mathbf{G}\mathbf{X}'\mathbf{Y}$$

(73)

$$q_j = \mathbf{b}^{0\prime}\mathbf{k}_j(\mathbf{k}_j'\mathbf{G}\mathbf{k}_j)^{-1}\mathbf{k}_j'\mathbf{b}^0 = \mathbf{Y}'\mathbf{X}\mathbf{G}'\mathbf{k}_j(\mathbf{k}_j'\mathbf{G}\mathbf{k}_j)^{-1}\mathbf{k}_j'\mathbf{G}\mathbf{X}'\mathbf{Y}$$

sînt numărătorul SS pentru verificarea ipotezei $\mathbf{k}_i'\boldsymbol{\beta} = 0$ și respectiv $\mathbf{k}_j'\boldsymbol{\beta} = 0$. Din problema 12 aceste sume de pătrate privite ca forme pătratice în $\mathbf{Y}$ care este presupusă a fi repartizate $N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, vor fi independente cînd :

$$\mathbf{X}\mathbf{G}'\mathbf{k}_i(\mathbf{k}_i'\mathbf{G}\mathbf{k}_i)^{-1}\mathbf{k}_i'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{k}_j(\mathbf{k}_j\mathbf{G}\mathbf{k}_j)^{-1}\mathbf{k}_j'\mathbf{G}\mathbf{X}' = \mathbf{0}.$$

O condiție suficientă pentru aceasta este

$$\mathbf{k}_i'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{k}_j = 0.$$

Deoarece $\mathbf{k}_j'\boldsymbol{\beta}$ este estimabilă, $\mathbf{k}_j' = \mathbf{t}_j'\mathbf{X}$ pentru anumiți $\mathbf{t}_j'$, și astfel condițiile devin

$$\mathbf{k}_i'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{X}'\mathbf{t}_j = \mathbf{k}_i'\mathbf{G}\mathbf{X}'\mathbf{t}_j = \mathbf{k}_i'\mathbf{G}\mathbf{k}_j = 0. \tag{74}$$

Astfel (74) este condiția care face q_i și q_j din (73) independente. Ea face de asemenea

$$(\mathbf{K}'\mathbf{G}\mathbf{K})^{-1} = \operatorname{diag}\{(\mathbf{k}_i'\mathbf{G}\mathbf{k}_i)^{-1}\} \quad \text{pentru} \quad i = 1, 2, \ldots, r$$

și astfel

$$Q = \sum_{i=1}^{r} \mathbf{b}^{0\prime}\mathbf{k}_i(\mathbf{k}_i'\mathbf{G}\mathbf{k}_i)^{-1}\mathbf{k}_i'\mathbf{b}^0 = \sum_{i=1}^{r} \frac{(\mathbf{k}_i'\mathbf{b}^0)^2}{\mathbf{k}_i'\mathbf{G}\mathbf{k}_i} = \sum_{i=1}^{r} q_i. \tag{75}$$

După (35), (74) este de asemenea condiția ca $\mathbf{k}_i'\mathbf{b}^0$ și $\mathbf{k}_j'\mathbf{b}^0$ să fie independente. Deci cînd $\mathbf{K}'\boldsymbol{\beta}$ constă în $r = r(\mathbf{X})$ funcții liniar independente, $\mathbf{k}_i'\boldsymbol{\beta}$ pentru $i = 1, 2, \ldots, r$ și cînd $\mathbf{k}_i'\mathbf{b}^0$ sînt independente, atunci numărătorul SS : Q pentru verificarea ipotezei $\mathbf{K}'\boldsymbol{\beta} = \mathbf{0}$ este după (72) nu numai egal cu SSR, ci și cu $\sum_{i=1}^{r} q_i$ pentru verificarea ipotezelor $\mathbf{k}_i'\boldsymbol{\beta} = \mathbf{0}$, pentru $i = 1, 2, \ldots, r$. Am demonstrat

TEOREMA 14. *Dacă, pentru* $i = 1, 2, \ldots, r$

$$\mathbf{k}_i' = \mathbf{k}_i'\mathbf{H}, \tag{76}$$

$$\mathbf{k}_i'\mathbf{G}\mathbf{k}_j = 0 \quad \textit{pentru} \quad i \neq j \tag{77}$$

$$\mathbf{k}_i' \textit{ sînt liniar independente,} \tag{78}$$

atunci

$F(H) = Q/r\,\hat{\sigma}^2$ *este statistica test pentru* $H : \mathbf{K}'\boldsymbol{\beta} = \mathbf{0}$

$F(H_i) = q_i/\hat{\sigma}^2$ *este statistica test pentru* $H_i : \mathbf{k}_i\boldsymbol{\beta} = 0$

și q_i, $i = 1, 2, \ldots, r$ *sînt independente cu*

$$q_i = \frac{(\mathbf{k}_i'\mathbf{b}^0)^2}{\mathbf{k}_i'\mathbf{G}\mathbf{k}_i}.$$

Scanned with OKEN Scanner

TEOREMA 15. *Dacă este adevărată ipoteza* $H : \mathbf{K}'\boldsymbol{\beta} = \mathbf{0}$

$[H : \mathbf{k}_i\boldsymbol{\beta} = 0]$ *atunci* $F(H) \sim F(r,\ n-r)[F(H_i) \sim F(1,\ n-r)]$.

Observație. Avem

$$\sqrt{\frac{q_i}{\hat{\sigma}^2}} = \frac{\mathbf{k}_i'\mathbf{b}^0}{\sqrt{\mathbf{k}_i'\mathbf{G}\mathbf{k}_i\hat{\sigma}^2}} \sim t(n-r)$$

statistică folosită pentru verificarea ipotezei H_i.

În cazul datelor „balansate" (date în care nmărul de observații este același în subclase) rezultatele anterioare conduc la mulțimea de valori pentru $\mathbf{k}_i'$ astfel încît $\mathbf{k}_i'\boldsymbol{\beta}$ sînt numite contraste ortogonale: *ortogonale* deoarece atunci $\mathbf{G}$ este astfel încît (77) se reduce la $\mathbf{k}_i'\mathbf{k}_j = 0$ și *contraste* deoarece $\mathbf{k}_i'\boldsymbol{\beta}$ poate fi exprimat ca suma diferențelor dintre elementele lui $\boldsymbol{\beta}$.

§ 5. Modele restrînse

5.1. Introducere

DEFINIȚIA 5. Relațiile ce există ca parte integrală a modelului, vor purta numele de restricții asupra modelului.

Observație. Restricțiile asupra modelului se deosebesc de relațiile ce se impun asupra soluțiilor ecuațiilor normale avînd drept scop simplificarea obținerii acestora; aceste relații vor purta numele de *constrîngeri asupra soluțiilor.*

DEFINIȚIA 6. Modelele care includ restricții poartă numele de *modele restrînse*

Modelele de care ne-am ocupat pînă acum poartă numele de *modele nerestrînse.*

Problema care se pune este să vedem cum metoda de estimare și de verificare a ipotezelor dezvoltată pentru modelele nerestrînse se aplică la modelele restrînse. În general considerăm ca parte a modelului mulțimea de restricții.

$$\mathbf{P}'\boldsymbol{\beta} = \boldsymbol{\delta}, \tag{79}$$

unde $\mathbf{P}'$ are rang maxim q(q = nr. de linii).
Atunci, modelul restrîns este $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ restrîns de $\mathbf{P}'\boldsymbol{\beta} = \boldsymbol{\delta}$.
La fel ca în (60) ajustînd modelul restrîns obținem

$$\begin{aligned} \mathbf{X}'\mathbf{X}\mathbf{b}_r^0 + \mathbf{P}'\boldsymbol{\theta} &= \mathbf{X}'\mathbf{X} \\ \mathbf{P}'\mathbf{b}_r^0 &= \boldsymbol{\delta} \end{aligned} \tag{80}$$

unde $2\boldsymbol{\theta}$ este multiplicatorul lui Lagrange și indicele r de la $\mathbf{b}_r^0$ arată că $\mathbf{b}_r^0$ este soluția ecuației normale în modelul restrîns. În rezolvarea lui (80) trebuie să distingem două cazuri după cum $\mathbf{P}'\boldsymbol{\beta}$ este estimabil sau nu, deoarece soluția nu este aceeași în ambele cazuri.

Scanned with OKEN Scanner

5.2. Restricții conținînd funcții estimabile

TEOREMA 16. *Dacă* $\mathbf{P}'\boldsymbol{\beta}$ *este estimabil,* (80) *are soluția*

$$\mathbf{b}_r^0 = \mathbf{b}^0 - \mathbf{GP}(\mathbf{P}'\mathbf{GP})^{-1}(\mathbf{P}'\mathbf{b}^0 - \boldsymbol{\delta}) \tag{81}$$

cu

$$M(\mathbf{b}_r^0) = \mathbf{H}\boldsymbol{\beta} - \mathbf{GP}(\mathbf{P}'\mathbf{GP})^{-1}(\mathbf{P}'\mathbf{H}\boldsymbol{\beta} - \boldsymbol{\delta}) = \mathbf{H}\boldsymbol{\beta}$$

$$\mathscr{C}(\mathbf{b}_r^0) = \mathscr{C}\{[\mathbf{I} - \mathbf{GP}(\mathbf{P}'\mathbf{GP})^{-1}\mathbf{P}']\mathbf{b}^0\} = \mathbf{G}[\mathbf{X}'\mathbf{X} - \mathbf{P}(\mathbf{P}'\mathbf{GP})^{-1}\mathbf{P}']\mathbf{G}'\sigma^2.$$

SS erorilor după ajustarea acestui model restrîns este

$$\mathrm{SSE}_r = (\mathbf{Y} - \mathbf{X}\mathbf{b}_r^0)'(\mathbf{Y} - \mathbf{X}\mathbf{b}_r^0)$$

și din (64) și (63)

$$\mathrm{SSE}_r = \mathrm{SSE} + (\mathbf{P}'\mathbf{b}^0 - \boldsymbol{\delta})'(\mathbf{P}'\mathbf{GP})^{-1}(\mathbf{P}'\mathbf{b}^0 - \boldsymbol{\delta}) \tag{82}$$

cu

$$M(\mathrm{SSE}_r) = (n-r)\sigma^2 + \mathbf{M}\mathbf{b}^{0\prime}\mathbf{P}(\mathbf{P}'\mathbf{GP})^{-1}\mathbf{P}'\mathbf{b}^0 - \boldsymbol{\delta}'(\mathbf{P}'\mathbf{GP})^{-1}\boldsymbol{\delta}.$$

Aplicînd problema 10 termenului din mijloc, folosind (4) și apoi (79)

$$M(\mathrm{SSE}_r) = (n - r + q)\sigma^2.$$

Deci am demonstrat

TEOREMA 17. *În modelul restrîns,*

$$\hat{\sigma}_r^2 = \frac{\mathrm{SSE}_r}{n-r+q}$$

este un estimator nedeplasat pentru dispersia erorii.

Observație. Ca indice, r se referă la modelul restrîns, iar în expresii $r = r(\mathbf{X})$.

$\mathbf{b}_r^0$ și SSE_r din (81) și (82) nefiind aceeași ca $\mathbf{b}^0$ și SSE indică că restricțiile de estimabilitate asupra parametrilor modelului afectează procesul de estimare. Totuși, acestea nu afectează estimabilitatea oricărei funcții care este estimabilă în modelul restrîns. Condiția de estimabilitate $M(\mathbf{t}'\mathbf{Y}) = \mathbf{k}'\boldsymbol{\beta}$ pentru anumiți $\mathbf{t}'$ rămîne neschimbată.

Totuși, ca funcție estimabilă, ea este o funcție de parametrii și prin urmare supusă la restricțiile $\mathbf{P}'\boldsymbol{\beta} = \boldsymbol{\delta}$. Acestea pot schimba forma lui $\mathbf{k}'\boldsymbol{\beta}$.

În general, funcția estimabilă $\mathbf{k}'\boldsymbol{\beta}$ se schimbă la $\mathbf{k}'\boldsymbol{\beta} + \boldsymbol{\lambda}'(\mathbf{P}'\boldsymbol{\beta} - \boldsymbol{\delta})$ unde, pentru ca aceasta să fie o funcție de $\mathbf{b}$, $\boldsymbol{\lambda}'$ trebuie să fie ales astfel încît $\boldsymbol{\lambda}'\boldsymbol{\delta} = \mathbf{0}$ (cînd $\boldsymbol{\delta}$ este nul, $\boldsymbol{\lambda}'$ poate fi orice vector). Atunci $\mathbf{k}'\boldsymbol{\beta}$ devine $\mathbf{k}'\boldsymbol{\beta} + \boldsymbol{\lambda}'\mathbf{P}'\boldsymbol{\beta} = (\mathbf{k}' + \boldsymbol{\lambda}'\mathbf{P}')\boldsymbol{\beta}$. Această funcție este de asemenea estimabilă în modelul nerestrîns deoarece sînt estimabile atît $\mathbf{k}'\boldsymbol{\beta}$ cît și $\mathbf{P}'\boldsymbol{\beta}$.

În modelul restrîns poate fi considerată ipoteza $H\colon \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ numai dacă este consistentă cu $\mathbf{P}'\boldsymbol{\beta} = \boldsymbol{\delta}$.

Scanned with OKEN Scanner

Cu această limitare a consistenței, ipoteza $\mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ este verificată în modelul restrîns considerînd modelul nerestrîns $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ supus atît restricțiilor $\mathbf{P}'\boldsymbol{\beta} = \boldsymbol{\delta}$ cît și ipotezele verificabile $\mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$. Modelul restrîns redus de ipoteza $\mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ poate fi numit modelul restrîns redus.

Punînd

$$\mathbf{Q}' = \begin{pmatrix} \mathbf{P}' \\ \mathbf{K}' \end{pmatrix} \text{ și } \ell = \begin{pmatrix} \boldsymbol{\delta} \\ \mathbf{m} \end{pmatrix}$$

minimizează $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ supus la

$$\mathbf{Q}'\boldsymbol{\beta} = \ell.$$

Deoarece atît $\mathbf{P}'$ cît și $\mathbf{K}'$ au rang maxim linie și liniile lor sînt mutual liniar independente, $\mathbf{Q}'$ are rang maxim linie și $\mathbf{Q}'\boldsymbol{\beta}$ este estimabilă. Minimizarea, conduce la soluția

$$\mathbf{b}^0_{r,H} = \mathbf{b}^0 - \mathbf{GQ}(\mathbf{Q}'\mathbf{GQ})^{-1}(\mathbf{Q}'\mathbf{b}^0 - \ell)$$

Suma pătratelor reziduale este:

$$\mathrm{SSE}_{r,H} = \mathrm{SSE} + (\mathbf{Q}'\mathbf{b}^0 - \ell)'(\mathbf{Q}'\mathbf{GQ})^{-1}(\mathbf{Q}'\mathbf{b}^0 - \ell)$$

și verificarea ipotezei $\mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ este bazată pe

$$(84) \qquad F(H_r) = (\mathbf{Q}'\mathbf{b}^0 - \ell)'(\mathbf{Q}'\mathbf{GQ})^{-1}(\mathbf{Qb}^0 - \ell)/s\hat{\sigma}^2_r \sim F(s, n - r + q)$$

unde $\hat{\sigma}^2_r$ este dat în (83)

Ca și în cazul funcțiilor estimabile, o ipoteză care este verificabilă în modelul nerestrîns este de asemenea verificabilă în modelul restrîns. Modificarea ipotezei de restricții poate schimba forma ipotezei, însă această modificarea a sa va fi verificabilă nu numai în modelul restrîns ci și în modelul nerestrîns

Deci are loc

TEOREMA 18. *Dacă* $\mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ *este verificabilă în modelul nerestrîns atunci pentru orice matrice* $\mathbf{L}(s \times q)$

$$(\mathbf{K}' + \mathbf{LP}')\boldsymbol{\beta} = \mathbf{m} + \mathbf{L}\boldsymbol{\delta}$$

va fi verificabilă în modelul restrîns; de asemenea va fi verificabilă în modelul nerestrîns.

5.3. Restricții conținînd funcții care nu sînt estimabile

Cînd restricțiile sînt $\mathbf{P}'\boldsymbol{\beta} = \boldsymbol{\delta}$ și $\mathbf{P}'\boldsymbol{\beta}$ nu este estimabilă, soluțiile lui (90) sînt, asemănător cu (69)

$$(85) \qquad \mathbf{b}^0_r = \mathbf{b}^0 + (\mathbf{H} - \mathbf{I})\mathbf{z}_1$$

unde $\mathbf{z}_1$, urmînd (67) satisface

$$(86) \qquad \mathbf{P}'(\mathbf{H} - \mathbf{I})\mathbf{z}_1 = \boldsymbol{\delta} - \mathbf{P}'\mathbf{GX}'\mathbf{Y}.$$

Scanned with OKEN Scanner

Deoarece $\mathbf{b}_r^0$ este o soluție a ecuațiilor normale $\mathbf{X'X\beta} = \mathbf{X'Y}$. Prin urmare, în acest caz

$$SSE_r = SSE$$

adică restricțiile nu afectează SS erorilor reziduală.

Ca mai înainte, includerea de restricții în model nu alterează estimabilitatea funcției care este estimabilă în modelul nerestrîns.

Există funcții care nu sînt estimabile în modelul nerestrîns însă pot fi estimabile în modelul restrîns. În general : „dacă $\mathbf{k'\beta}$ este estimabilă în modelul nerestrîns atunci $\mathbf{k'\beta} + \boldsymbol{\lambda}'(\mathbf{P'\beta} - \boldsymbol{\delta})$ este estimabilă în modelul restrîns ; pentru a elimina $\boldsymbol{\delta}$ cînd este nenul, $\boldsymbol{\lambda}'$ trebuie să fie astfel încît $\boldsymbol{\lambda}'\boldsymbol{\delta} = \mathbf{0}$; și funcția $\mathbf{K'\beta} + \boldsymbol{\lambda}'\mathbf{P'\beta}$ este estimabilă în modelul restrîns".

Pentru verificareai ipotezei $\mathbf{K'\beta} = \mathbf{m}$.

$$SSE_{r,H} = SSE_H$$

și deci statistica F pentru verificarea ipotezei este identică cu aceea pentru modelul nerestrîns.

SSE și SSE_H sînt calculate în mod obișnuit. Astfel statistica F este calculată ca în (59)

Astfel ipotezele care sînt verificabile în modelul nerestrîns sînt de asemenea verificabile în modelul restrîns ; însă aplicarea restricțiilor poate schimba forma lor astfel că cu toate că ele sînt verificabile în modelul restrîns ele pot să nu fie verificabile în modelul nerestrîns. În general : „dacă $\mathbf{K'\beta} = \mathbf{m}$ este verificabilă în modelul nerestrîns atunci pentru orice matrice $\mathbf{L}(s \times q)$, $(\mathbf{K'} + \mathbf{LP'})\boldsymbol{\beta} = \mathbf{m} + \mathbf{L}\boldsymbol{\delta}$ este verificabilă în modelul restrîns : nu va fi verificabilă în modelul nerestrîns".

PROBLEME

la capitolele XXXI și XXXII

Definiția 1. Ecuațiile liniare $\mathbf{AX=Y}$ sînt definite ca fiind consistente dacă orice relații liniare ar exista printe liniile lui **A** acestea există de asemenea printre elementele corespunzătoare ale lui **Y**.

1. Să se arate că ecuația consistentă $\mathbf{AX} = \mathbf{Y}$ are o soluție $\mathbf{X} = \mathbf{GY}$ dacă și numai dacă $\mathbf{AGA} = \mathbf{A}$.

Definiția 2. Inversa generalizată a matricii **A** este matricea **G** care satisface relația

$$\mathbf{AGA} = \mathbf{A}.$$

2. Să se arate că dacă **A** are q coloane și dacă **G** este inversa generalizată a lui **A**, atunci ecuațiile consistente $\mathbf{AX} = \mathbf{Y}$ au soluția

$$\tilde{\mathbf{X}} = \mathbf{GY} + (\mathbf{GA} - \mathbf{I})\mathbf{z}$$

unde **z** este orice vector arbitrar de ordin q.

3. Să se arate că pentru orice **G** specificat, toate soluțiile ecuațiilor consistente $\mathbf{AX} = \mathbf{Y}$, sînt generate de

$$\tilde{\mathbf{X}} = \mathbf{GY} + (\mathbf{GA} - \mathbf{I})\mathbf{z}$$

pentru **z** arbitrar.

4. Să se arate că dacă **A** este o matrice cu q coloane și de rang r și dacă **Y** este un vector nenul, atunci $q - r + 1$ este numărul de soluții liniar independente ale ecuațiilor consistente $\mathbf{AX} = \mathbf{Y}$.

Scanned with OKEN Scanner

5. Să se arate că valoarea lui $\mathbf{k'\tilde{X}}$ este invariantă în raport cu soluția $\tilde{\mathbf{X}}$ a ecuației $\mathbf{AX = Y}$ dacă $\mathbf{k'H = k'}$, unde $\mathbf{H = GA}$ și $\mathbf{AGA = A}$.

6. Să se arate că dacă $\mathbf{G}$ este inversa generalizată a lui $\mathbf{X'X}$, atunci

i) $\mathbf{G'}$ este de asemenea inversa generalizată a lui $\mathbf{X'X}$

ii) $\mathbf{XGX'X = X}$, adică $\mathbf{GX'}$ este inversa generalizată a lui $\mathbf{X}$

iii) $\mathbf{XGX'}$ este invariantă sub $\mathbf{G}$

iv) $\mathbf{XGX'}$ este simetrică, indiferent dacă $\mathbf{G}$ este sau nu.

7. Să se arate că pentru toate inversele generalizate $\mathbf{G}$ ale lui $\mathbf{A}$, $\tilde{\mathbf{x}} = \mathbf{Gy}$ generează toate soluțiile ecuațiilor consistente $\mathbf{Ax = y}$.

8. Să se arate că $\mathbf{A'A}$ este pozitiv definită cînd A este de rang maxim coloană și este semipozitiv definită în rest.

9. Să se arate că $\mathbf{AA'}$ este pozitiv definită cînd A este de rang maxim linie și semipozitiv definită în rest.

10. Dacă $\mathbf{X} \sim \mathbf{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ să se arate că

i) $M(\mathbf{X'AX}) = \mathrm{Ur}(\mathbf{A\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$ (adevărată și cînd $\mathbf{X}$ nu este normal)

ii) cumulantul de ordinul r al formei $\mathbf{X'AX}$ este $\mathcal{K}_r(\mathbf{X'AX}) = 2^{r-1}(r-1)![\mathrm{Ur}(\mathbf{A\Sigma})^r + r\boldsymbol{\mu}'\mathbf{A}(\boldsymbol{\Sigma}\mathbf{A})^{r-1}\boldsymbol{\mu}]$ și

iii) covarianța dintre $\mathbf{X}$ și $\mathbf{X'AX}$ este

$$\mathrm{cov}\,(\mathbf{X}, \mathbf{X'AX}) = 2\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu}.$$

11. Să se arate că dacă $\mathbf{X} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ atunci $\mathbf{X'AX} \sim \chi^2\left[r(\mathbf{A}), \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}\right]$ dacă și numai dacă

$$\mathbf{A\Sigma} \text{ este idempotentă.}$$

12. Să se arate că dacă $\mathbf{X} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ atunci $\mathbf{X'AX}$ și $\mathbf{BX}$ sînt independente dacă și numai dacă $\mathbf{B\Sigma A = 0}$

13. Să se arate că dacă $\mathbf{X} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, formele pătratice $\mathbf{X'AX}$ și $\mathbf{X'BX}$ sînt independente dacă și numai dacă

$$\mathbf{A\Sigma B = B\Sigma A = 0}$$

14. Fie $\mathbf{X} \sim \mathbf{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{A}_i$ o matrice $n \times n$ simetrică de rang k_i, $i = 1, \ldots, p$ și $\mathbf{A} = \sum_{i=1}^{n} \mathbf{A}_i$ care este simetrică, cu rangul k. Să se arate că

a) $\mathbf{X'A}_i\mathbf{X} \sim \chi^2\left(k_i, \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}_i\boldsymbol{\mu}\right)$.

b) $\mathbf{X'A}_i\mathbf{X}$ sînt independente

c) $\mathbf{X'AX} \sim \chi^2\left(k, \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}\right)$

dacă și numai dacă

I) oricare două din condițiile

i) $\mathbf{A}_i\boldsymbol{\Sigma}$ idempotentă pentru toți i

ii) $\mathbf{A}_i\boldsymbol{\Sigma}\mathbf{A}_j = 0$ pentru toți $i < j$

iii) $\mathbf{A\Sigma}$ idempotentă

sînt satisfăcute

sau II) este adevărată iii) și iv) $k = \sum_{i=1}^{p} k_i$

sau III) este adevărată iii) și v) $\mathbf{A}_1\boldsymbol{\Sigma}, \ldots, \mathbf{A}_{(p-1)}\boldsymbol{\Sigma}$ sînt idempotente și $\mathbf{A}_p\boldsymbol{\Sigma}$ este nenegativ definită.

Scanned with OKEN Scanner

Capitolul XXXIII

ANALIZA REGRESIEI PRIN RĂDĂCINI CARACTERISTICE

§ 1. Introducere

O atenţie deosebită se acordă metodelor alternative de estimare a parametrilor în modelul regresiei liniare multiple cînd matricea variabilelor independente este rău condiţionată.

Să considerăm modelul

$$\mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

unde : $\mathbf{Y}$ este vectorul variabilelor aleatoare (răspuns) observate $(n \times 1)$,
β_0 şi $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_k)$ sînt parametrii necunoscuţi,
$\mathbf{1}$ este vectorul $(n \times 1)$ unitate,
$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k]$ este o matrice $(n \times k)$ de rang k de variabile independente cunoscute, standardizate astfel încît :

$$\sum_{i=1}^{n} X_{ij} = \mathbf{X}_j'\mathbf{1} = 0, \quad \sum_{i=1}^{n} X_{ij}^2 = \mathbf{X}_j'\mathbf{X}_j = 1, \; j = 1, \ldots, k,$$

$\boldsymbol{\varepsilon}$ este un vector $n \times 1$ de variabile aleatoare necorelate $N(0, \sigma^2)$, neobservabile.

Dorim să găsim un estimator al lui (1) de forma

$$\mathbf{Y} = b_0 \mathbf{1} + \mathbf{X}\mathbf{b} \tag{2}$$

unde b_0 şi $\mathbf{b}_1$ sînt estimatorii lui β_0 şi $\boldsymbol{\beta}$. Deoarece matricea $\mathbf{X}$ conţine variabile independente standardizate, luăm $b_0 = \bar{Y}$. Considerăm estimatorul de C.M.M.P. pentru $\boldsymbol{\beta}$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Avem

$$M(b_j) = \beta_j, \qquad D^2(b_j) = c_{jj}\sigma^2$$

unde $(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{C}$ şi c_{jj} este elementul diagonal al matricii $(\mathbf{X}'\mathbf{X})^{-1}$. Estimatorii de C.M.M.P. sînt nedeplasaţi şi de dispersie minimă pentru parametrii modelului (1), în presupunerea că acesta este corect.

Scanned with OKEN Scanner

În cazul în care **X** este rău condiționată se obțin estimatori cu performanțe slabe. Pentru a vedea efectul rău condiționării asupra estimatorilor de C.M.M.P. partiționăm **X** astfel încît

$$\mathbf{X} = [\mathbf{x}_j : \mathbf{X}_j^*],$$

unde $\mathbf{x}_j$ este coloana observațiilor asupra variabilei independente j și $\mathbf{X}_j^*$ este matricea observațiilor asupra celorlalte $k-1$ variabile independente rămase.
Atunci

$$c_{jj} = [\mathbf{x}_j'\mathbf{x}_j - \mathbf{x}_j'\mathbf{X}_j^*(\mathbf{X}_j^{*'}\mathbf{X}_j^*)^{-1}\mathbf{X}_j^{*'}\mathbf{x}_j]^{-1} = [1 - R_j^2]^{-1},$$

unde R_j este coeficientul de corelație multiplu al variabilei independente j regresat asupra celorlalte $k-1$ variabile independente rămase. Astfel :

$$D^2(b_j) = \sigma^2/(1-R_j^2).$$

Dacă $\mathbf{x}_j$ este ortogonal cu $\mathbf{X}_j^*$, atunci $R_j^2 = 0$. Pe de altă parte, dacă există o dependență liniară între $\mathbf{x}_j$ și unele din coloanele lui $\mathbf{X}_j^*$, R_j^2 poate fi foarte apropiat de 1. Ne vom referi la acest tip de rău condiționare ca la aproape singularitate. Aceasta apare frecvent în practică și deoarece $D^2(b_j)$ este mare, probabilitatea de a obține un estimator cu performanțe slabe poate fi foarte înaltă cînd **X** este aproape singulară. O formă de estimare deplasată a parametrilor folosește regresia ridge.

Webster, Gunst, Mason (1974) modifică procedeul C.M.M.P. folosind rădăcinile proprii și vectorii proprii ai matricei de corelație a variabilelor dependente și independente. Acest procedeu :

i) identifică aproape singularitatea matricei **X**
ii) determină dacă aproape singularitățile au valoare predictivă ;
iii) obține estimații de C.M.M.P. modificate pentru parametrii care ajustează aproape singularitățile nepredictive.

§ 2. Obținerea ecuațiilor de predicție

Considerăm din nou modelul (1) și fie

$$Y_i^* = (Y_i - \bar{Y})/\eta \quad \text{unde} \quad \eta^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2,$$

Definim matricea

$$\mathbf{A} = [\mathbf{Y}^* : \mathbf{X}],$$

adică matricea $n \times (k+1)$ a variabilelor dependente standardizate și a variabilelor independente. Astfel $\mathbf{A}'\mathbf{A}$ este *matricea de corelație* a variabilelor dependente și independente și are rădăcinile și vectorii proprii definiți de

$$|\mathbf{A}'\mathbf{A} - \lambda_j\mathbf{I}| = 0$$

Scanned with OKEN Scanner

şi

$$(\mathbf{A}'\mathbf{A} - \lambda_j \mathbf{I})\gamma_j = \mathbf{0}, \quad j = 0, 1, \ldots, k.$$

Notăm elementele vectorului propriu j prin

$$\gamma_j' = (\gamma_{0j}, \gamma_{1j}, \ldots, \gamma_{kj})$$

şi fie

$$\gamma_j^{0\prime} = (\gamma_{1j}, \ldots, \gamma_{kj}),$$

adică γ_j^0 conţine toate elementele lui γ_j, exceptîndu-l pe primul. Fie

$$\Gamma = (\gamma_0 \gamma_1 \ldots \gamma_k) \quad \text{şi} \quad \Lambda = \operatorname{diag}(\lambda_0, \lambda_1, \ldots, \lambda_k).$$

Atunci

$$\Gamma'(\mathbf{A}'\mathbf{A})\Gamma = \Lambda \quad \text{şi} \quad \mathbf{A}'\mathbf{A} = \Gamma\Lambda\Gamma'.$$

avem

$$\mathbf{A}\gamma_j = \begin{bmatrix} Y_1^* \gamma_{0j} + \sum_{r=1}^{k} X_{1r}\gamma_{rj} \\ Y_2^* \gamma_{0j} + \sum_{r=1}^{k} X_{2r}\gamma_{rj} \\ \cdots\cdots\cdots\cdots \\ Y_n^* \gamma_{0j} + \sum_{r=1}^{k} X_{nr}\gamma_{rj} \end{bmatrix} \tag{3}$$

şi

$$\lambda_j = (\mathbf{A}\gamma_j)'(\mathbf{A}\gamma_j) = \gamma_j'(\mathbf{A}'\mathbf{A})\gamma_j =$$

$$= \sum_{i=1}^{n} \left(Y_i^* \gamma_{0j} + \sum_{r=1}^{k} X_{ir}\gamma_{rj} \right)^2. \tag{4}$$

Dacă orice $\lambda_j = 0$, printre unele sau toate coloanele lui $\mathbf{A}$ există o dependenţă liniară exactă. Dacă γ_{0j} corespunzător este $\neq 0$ există predictorul exact:

$$\hat{Y}_i = \bar{Y} - \eta\gamma_{0j}^{-1} \sum_{r=1}^{k} X_{ir}\gamma_{rj}.$$

Dacă $\lambda_j = 0$ şi $\gamma_{0j} = 0$ printre coloanele lui $\mathbf{X}$ există o dependenţă liniară exactă, relaţia fiind

$$\sum_{r=1}^{k} X_{ir}\gamma_{rj} = 0.$$

În general nici una din rădăcinile proprii nu va fi zero, însă unele pot fi foarte mici. Rădăcinile proprii cu valori mici însă nenule indică aproape singularitatea.

Din (3) se pot defini (cu condiţia că toţi $\gamma_{0j} \neq 0$) $k+1$ ecuaţii de predicţie

$$\hat{\mathbf{Y}}_j = \bar{Y}\mathbf{1} - \eta\gamma_{0j}^{-1}\mathbf{X}\boldsymbol{\gamma}_j^0, \quad j = 0, 1, \ldots, k \tag{5}$$

care sînt de forma (2) cu

$$\mathbf{b} = -\eta\gamma_{0j}^{-1}\boldsymbol{\gamma}_j^0.$$

Dacă ecuaţia de predicţie j este folosită numai pentru predicarea valorilor variabilei dependente se verifică că suma pătratelor rezidurilor pentru acest predictor este

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_{ij})^2 = (\mathbf{Y} - \hat{\mathbf{Y}}_j)'(\mathbf{Y} - \hat{\mathbf{Y}}_j) = \eta^2\lambda_j^*$$

unde $\lambda_j^* = \lambda_j/\gamma_{0j}^2$.

Nici una dintre ecuaţiile individuale de predicţie (5) nu va fi ea însăşi un predictor bun. Prin urmare, pentru a obţine estimaţii pentru parametrii modelului vor fi folosite combinaţii liniare de aceşti predictori. Astfel considerăm următoarea combinaţie liniară arbitrară de predictori de tip (5)

$$\hat{\mathbf{Y}} = \sum_{j=0}^{k} a_j\gamma_{0j}\hat{\mathbf{Y}}_j. \tag{6}$$

Impunînd restricţia

$$\sum_{j=0}^{k} a_j\gamma_{0j} = 1,$$

obţinem

$$\hat{\mathbf{Y}} = \bar{Y}\cdot\mathbf{1} - \eta\mathbf{X}\left(\sum_{j=0}^{k} a_j\boldsymbol{\gamma}_j^0\right) \tag{7}$$

care este din nou de forma (2) cu $\mathbf{b} = -\eta\sum_{j=0}^{k} a_j\boldsymbol{\gamma}_j^0$.

Folosind acest predictor suma pătratelor rezidurilor este

$$(\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = \eta^2\mathbf{a}'\Lambda\,\mathbf{a} = \eta^2\sum_{j=0}^{k} a_j^2\lambda_j$$

unde $\mathbf{a} = (a_0 a_1 \ldots a_k)$.

§ 3. Interpretarea geometrică a rădăcinilor şi vectorilor proprii ai matricei A'A

O problemă care nu trebuie trecută cu vederea cînd discutăm aproape singularitatea lui $\mathbf{A}'\mathbf{A}$ este dacă aceste aproape singularităţi conţin informaţii asupra modelului (1). Pentru a răspunde la această întrebare considerăm cele n puncte $(Y_i^*, X_{i1}, X_{i2}, \ldots, X_{ik})$, $i = 1, 2, \ldots, n$ corespunzătoare datelor în spţiul Euclidian $k+1$ dimensional definit de axele mutual

Scanned with OKEN Scanner

ortogonale $Y^*, X_1, X_2, \ldots, X_k$. Vectorii proprii ai matricei $\mathbf{A'A}$ definesc o a doua mulțime de axe mutual ortogonale $Z_0, Z_1, Z_2, \ldots, Z_k$, unde Z_j este axa definită de γ_j, $j = 0, 1, \ldots, k$. Direcția axei Z_j relativă la originea axelor este dată de vectorul sumă $\sum_{r=0}^{k} \gamma_{rj}\mathbf{e}_r$ unde $\mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_k$ sînt vectorii de lungime unitate din origine în direcția axelor $Y^*, X_1, \ldots, X_k$. Primul element al lui γ_j reprezintă cosinusul unghiului dintre axele Y^* și Z_j în timp ce γ_{rj} $(r = 1, 2, \ldots, k)$ reprezintă cosinusul unghiului dintre axele X_r și Z_j. Presupunînd că vectorii proprii sînt normalizați, γ_j sînt unic determinați, exceptînd constanta de multiplicare -1.

Rădăcina proprie corespunzătoare la un vector propriu particular măsoară răspîndirea celor n puncte în direcția definită de vectorul propriu. Cu alte cuvinte, λ_j este suma pătratelor proiecțiilor celor n puncte pe axa Z_j. O valoare mică a lui λ_j indică că există o variabilitate mică în direcția Z_j, adică

$$Z_{ij} = Y_i^* \gamma_{0j} + \sum_{r=1}^{k} X_{ir}\gamma_{rj}$$

este apropiat de zero pentru $i = 1, 2, \ldots, n$. Dacă γ_{0j} este apropiat de zero axa Z_j este aproape ortogonală cu axa Y^*. Deci dacă atît λ_j cît și γ_{0j} sînt mici, vectorul propriu γ_j indică o dependență liniară puternică numai între variabilele independente care produc schimbări mici sau deloc în variabila dependentă.

Estimatorul de C.M.M.P. este o combinație liniară de toți cei $k+1$ vectori proprii, incluzînd vectorii proprii ce corespund la singularități aproape nepredictive. *Webster, Gunst, Mason* (1974) modifică estimatorul de C.M.M.P. folosind numai combinațiile liniare care nu au atît pe λ_j cît și pe γ_{0j} mici. În acest mod estimațiile coeficienților de regresie sînt ajustați pentru efectul de aproape singularitate și nepredictiv.

§ 4. Estimatori de C.M.M.P. și estimatori de C.M.M.P. modificați

Folosind combinația liniară (6) cu $\sum_{j=0}^{k} a_j\gamma_{0j} = 1$ suma patratelor reziduriilor este dată de

$$(\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = \eta^2 \mathbf{a}' \Lambda \mathbf{a}.$$

Dacă $\mathbf{a}$ este ales să minimizeze suma pătratelor rezidurilor, (6) dă predictorul de C.M.M.P. obișnuit. Astfel dorim să minimizăm

$$f(a) = \eta^2 \sum_{j=0}^{k} a_j^2 \lambda_j - 2\mu_0 \left(\sum_{j=0}^{k} a_j \gamma_{0j} - 1 \right), \tag{8}$$

unde $-2\mu_0$ este multiplicatorul lui Lagrange. Soluția este

$$a_j = \eta^{-2} \lambda_j^{-1} \gamma_{0j} \mu_0, \quad j = 0, 1, \ldots, k.$$

Scanned with OKEN Scanner

Deoarece

$$\sum_{j=0}^{k} a_j \gamma_{0j} = 1, \quad \mu_0 = \eta^2 \left(\sum_{j=0}^{k} \lambda_j^{*-1} \right)^{-1}$$

unde $\lambda_j^* = \lambda_j / \gamma_{0j}^2$. Deci

$$a_j = \gamma_{0j} \lambda_j^{-1} \left(\sum_{l=0}^{k} \lambda_l^{*-1} \right)^{-1}, \quad j = 0, 1, \ldots, k. \tag{9}$$

Din (6) coeficienţii de C.M.M.P. sînt

$$\mathbf{b} = -\eta \left(\sum_{l=0}^{k} \lambda_l^{*-1} \right)^{-1} \sum_{j=0}^{k} \gamma_{0j} \lambda_j^{-1} \gamma_j^0 \tag{10}$$

cu suma pătratelor rezidurilor

$$\mathrm{SSE}(X_1, X_2, \ldots, X_k) = \eta^2 \left(\sum_{l=0}^{k} \lambda_l^{*-1} \right)^{-1} = \mu_0. \tag{11}$$

Presupunem că vectorii proprii $\gamma_0, \gamma_1, \ldots, \gamma_{p-1}$ corespund la aproape singularităţile nepredictive. Estimatorul de C.M.M.P. anterior poate fi ajustat luînd $a_0 = a_1 = \ldots = a_{p-1} = 0$. Atunci minimizînd (8), obţinem

$$a_j = \gamma_{0j} \lambda_j^{-1} \left(\sum_{l=p}^{k} \lambda_l^{*-1} \right)^{-1}, \quad j = p, p+1, \ldots, k \tag{12}$$

astfel că coeficienţii de C.M.M.P. modificaţi sînt

$$\mathbf{b}^* = -\eta \left(\sum_{l=p}^{k} \lambda_l^{*-1} \right)^{-1} \sum_{j=p}^{k} \gamma_{0j} \lambda_j^{-1} \gamma_j^0 \tag{13}$$

cu suma pătratelor rezidurilor

$$\mathrm{SSE}^*(X_1, X_2, \ldots, X_k) = \eta^2 \left(\sum_{l=p}^{k} \lambda_l^{*-1} \right)^{-1}. \tag{14}$$

Observaţie. Dacă $\mathbf{X}$ nu este de rang maxim coloană, adică $\mathbf{X}'\mathbf{X}$ este singulară, acest procedeu poate fi aplicat şi minimizatea lui (8) dă rezultate identice cu (12) la (14). Aceasta urmează din faptul că o matrice $\mathbf{X}'\mathbf{X}$ singulară implică că unele dintre valorile proprii λ_j şi corespunzători γ_{0j} ai matricii $\mathbf{A}'\mathbf{A}$ vor fi zero, echivalent cu a lua a_j din (8) apropiat de zero. Deci din acest procedeu pot fi obţinute soluţiile ecuaţiilor normale indiferent dacă $\mathbf{X}$ este de rang maxim coloană.

Estimaţiile obţinute din ecuaţiile (10) şi (13) sînt adeseori diferite cînd $\mathbf{X}$ este aproape singulară. Motivul acestui fapt este că a_j corespunzători la vectorii proprii ce relevă aproape singularităţi nepredictive sînt adeseori mari relativ la ceilalţi a_j rămaşi. Cînd se realizează aceasta termenii $a_j \gamma_j^0$, $j = 0, 1, \ldots, p-1$ pot domina $\mathbf{b}$. Înlăturînd aceşti termeni, obţinem o estimaţie mai bună pentru parametrul $\boldsymbol{\beta}$.

Scanned with OKEN Scanner

O altă măsură a efectului procedeului modificat este suma pătratelor rezidurilor. Ecuaţia (11) este suma pătratelor rezidurilor folosind C.M.M.P. în timp ce (14) este pentru procedeul modificat. Prin definiţie (14) este întotdeauna mai mare decît (11). Dacă se înlătură aproape singularităţile ne-predictive din estimaţii atunci (14) nu va diferi puternic de (11). Alternativ, estimaţiile lui σ^2 folosind (11) şi (14) vor fi apropiate.

§ 5. Eliminarea variabilelor independente

În analiza datelor o problemă de mare interes o prezintă alegerea *celei mai bune* submulţimi de k variabile independente pentru a fi folosită în ecuaţia de predicţie. O soluţie a acestei probleme ar fi examinarea tuturor celor 2^k regresii posibile ale variabilelor dependente asupra submulţimilor de variabile independente şi de a aplica un criteriu pentru determinarea celor mai buni predictori. Această soluţie are două neajunsuri
i) restricţii economice, de timp şi de calcul şi
iii) definiţia *celei mai bune*

Cu mai mult de 10 variabile independente i) face *toate regresiile posibile* o abordare impracticabilă.

Presupunem că dorim să eliminăm din predictor o mulţime specifică de s variabile independente. Printr-o alegere convenabilă a lui **a** din (6) putem face unii coeficienţi din **b** să fie zero, în felul acesta eliminînd variabilele independente corespunzătoare. Pentru uşurinţă presupunem înlăturate primele s variabile independente. Din (6) şi (7) aceasta cere ca să alegem **a** astfel încît să minimizeze $\eta^2\mathbf{a}'\wedge\mathbf{a}$ supusă la restricţiile

$$(15)\qquad \begin{cases} 1)\ \sum_{j=0}^{k} a_j\gamma_{0j} = 1 \\ 2)\ \sum_{j=0}^{k} a_j\gamma_{rj} = 0, \qquad r = 1, 2, \ldots, s. \end{cases}$$

Folosind multiplicatorii lui Lagrange $\{-2\mu_r\,;\ r = 0, 1, \ldots, s\}$ **a** este ales să minimizeze

$$f(\mathbf{a}) = \eta^2 \sum_{j=0}^{k} a_j^2\lambda_j - 2\mu_0\left(\sum_{j=0}^{k} a_j\gamma_{0j} - 1\right) - 2\sum_{r=1}^{s}\mu_r\left(\sum_{j=0}^{k} a_j\gamma_{rj} - 0\right).$$

Minimizînd $f(\mathbf{a})$ şi impunînd restricţiile (15), rezultă următorul sistem de $s+1$ ecuaţii :

$$(16)\qquad \begin{aligned} \eta^2 &= \rho_{00}\mu_0 + \rho_{01}\mu_1 + \ldots + \rho_{0s}\mu_s \\ 0 &= \rho_{10}\mu_0 + \rho_{11}\mu_1 + \ldots + \rho_{1s}\mu_s \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ 0 &= \rho_{s0}\mu_0 + \rho_{s1}\mu_1 + \ldots + \rho_{ss}\mu_s, \end{aligned}$$

Scanned with OKEN Scanner

unde

$$\rho_{uv} = \rho_{vu} = \sum_{j=0}^{k} \gamma_{uj}\gamma_{vj}\lambda_j^{-1}.$$

Notăm că $\rho_{00} = \sum_{j=0}^{k} \lambda_j^{*-1}$ și prin urmare

$$\mathrm{SSE}(X_1, X_2, \ldots, X_k) = \eta^2 \rho_{00}^{-1}.$$

Folosind soluțiile pentru a_j (în termeni de μ_r) obținute în minimizarea lui $f(\mathbf{a})$ și retricțiile (15) se verifică că suma patratelor reziduriilor înlăturînd aceste s variabile independente este soluția lui (16) pentru μ_0, adică

$$\mathrm{SSE}(X_{s+1}, X_{s+2}, \ldots, X_k) = \eta^2 \sum_{j=0}^{k} a_j^2 \lambda_j = \mu_0.$$

Este important de notat că a_j nu trebuiesc determinate în cazul în care se urmărește numai dacă acest model redus este corespunzător.

Statistica F folosită pentru compararea modelului redus cu modelul maxim este

$$F = \left[\frac{\mathrm{SSE}(X_{s+1}, X_{s+2}, \ldots, X_k)}{\mathrm{SSE}(X_1, X_2, \ldots, X_k)} - 1\right] \cdot \frac{n-k-1}{s} =$$

$$= (\mu_0(\eta^2\rho_{00})^{-1} - 1)(n-k-1)/s. \tag{17}$$

De îndată ce modelul redus a fost determinat pot fi obținuți μ_r rămași și a_j din

$$a_j = \eta^{-2}\lambda_j^{-1} \sum_{r=0}^{s} \mu_r \gamma_{rj}, \qquad j = 0, 1, \ldots, k.$$

Introducînd a_j în (6) sau (7) obținem predictorul de C.M.M.P. obișnuit pentru modelul redus. Poate fi dezvoltată o tehnică alternativă pentru procedeul modificat al C.M.M.P. Se modifică ρ_{uv} în procedeul C.M.M.P. Fie $\gamma_0, \gamma_1, \ldots, \gamma_{p-1}$ vectorii proprii ce corespund la aproape singularitățile ne-predictive. ρ_{uv} din (16) sînt înlocuiți prin

$$\rho_{uv}^{*} = \sum_{j=p}^{k} \gamma_{uj}\gamma_{vj}\lambda_j^{-1}.$$

Atunci

$$\mathrm{SSE}^{*}(X_{s+1}, X_{s+2}, \ldots, X_k) = \eta^2 \sum_{j=p}^{k} a_j^2 \lambda_j.$$

este din nou soluția lui (16) pentru μ_0. Coeficienții de C.M.M.P. modificați înlăturînd $X_1, X_2, \ldots, X_s$ din predictor sînt dați de

$$\mathbf{b}^{*} = -\eta \sum_{j=0}^{k} a_j \gamma_j^0,$$

Scanned with OKEN Scanner

unde

$$a_j = \begin{cases} 0, & j = 0, 1, \ldots, p-1 \\ \eta^{-2}\lambda_j^{-1} \sum_{r=0}^{s} \mu_r \gamma_{rj}, & j = p, p+1, \ldots, k. \end{cases}$$

Pentru a determina care dintre variabilele independente trebuie considerată prima pentru a fi înlăturată, formăm pentru fiecare $r = 1, 2, \ldots, k$ din (16) următoarele două ecuații

$$\eta_i^2 = \rho_{00}\mu_0 + \rho_{0r}\mu_r \tag{18}$$

$$0 = \rho_{r0}\mu_0 + \rho_{rr}\mu_r$$

și rezolvăm pentru μ_0 în fiecare caz. Atunci

$$\text{SSE}(\textit{toți } X, \textit{ exceptînd } X_r) = \mu_0.$$

Cea mai mică dintre cele k valori ale lui μ_0 determină prima variabilă independentă ce trebuie înlăturată. Testul este analog cu (17), adică statistica F este

$$F = \left[\frac{\text{SSE}(\textit{toți } X, \textit{ exceptînd } X_{(1)})}{\eta_i^2 \rho_{00}}\right](n-k-1), \tag{19}$$

unde SSE $(\textit{toți } X \textit{ exceptînd } X_{(1)}) = \min_r\{\text{SSE } (\textit{toți } X \textit{ exceptînd } r)\}$. Pasul următor constă în a considera $X_{(1)}$ înlăturat și o altă variabilă independentă $X_{(2)}$ astfel încît

$$\text{SSE } (\textit{toți } X, \textit{ exceptînd } X_{(1)} \textit{ și } X_{(2)}) =$$
$$= \min_{r \neq (1)} \{\text{SSE } (\textit{toți } X, \textit{ exceptînd } X_{(1)} \textit{ și } X_r)\}.$$

Se continuă în acest mod pînă cînd statistica F depășește valoarea critică.

§ 6. Proprietăți asimptotice

Obținem proprietățile asimptotice ale estimatorilor coeficienților de regresie în termeni de variabilele predictor (nestandardizate), astfel că scriem modelul (1) ca

$$\mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \tag{20}$$

unde păstrăm presupunerea că variabilele predictor din (20) sînt centrate, adică $\mathbf{Z}'\mathbf{1} = \mathbf{0}$. Relația dintre cantitățile din (1) și (20) este că

$$X_j = (Z_j'Z_j)^{-1/2}Z_j \text{ și } \beta_j = (Z_j'Z_j)^{1/2}\alpha_j, \quad j = 1, 2, \ldots, k$$

Scanned with OKEN Scanner

sau echivalent că

$$\mathbf{X} = \mathbf{Z}\mathbf{D}_{\mathbf{Z}}^{-1/2} \quad \text{şi} \quad \boldsymbol{\beta} = \mathbf{D}_{\mathbf{Z}}^{1/2}\boldsymbol{\alpha}$$

unde

$$\mathbf{D}_{\mathbf{Z}} = \operatorname{diag}(Z_1'Z_1, \ldots, Z_k'Z_k).$$

Pentru a determina forma limită a lui $\mathbf{A}'\mathbf{A}$ mai presupunem că

$$n^{-1}\mathbf{Z}'\mathbf{Z} \xrightarrow[n\to\infty]{} \mathbf{Q}, \tag{21}$$

$\mathbf{Q}$ fiind o matrice simetrică, pozitiv definită.

Urmează că

$$\mathbf{X}'\mathbf{X} \to \mathbf{R} = \mathbf{D}_{\mathbf{Q}}^{-1/2}\mathbf{Q}\mathbf{D}_{\mathbf{Q}}^{-1/2}$$

unde $\mathbf{D}_{\mathbf{Q}} = \operatorname{diag}(q_{11}, \ldots, q_{kk})$ şi q_{11} este elementul diagonal j al lui $\mathbf{Q}$.

Presupunerea (21) este o condiţie suficientă pentru consistenţa şi normalitatea asimptotică (cînd pentru ε_i nu este presupusă nici o formă pentru repartiţie) a estimatorului de C.M.M.P. pentru $\boldsymbol{\alpha}$. White, Gunst (1979) consideră forma limită pentru $\mathbf{A}'\mathbf{A}$ bazîndu-se pe $\mathbf{Z}$ fixat. În presupunerea (21)

$$p\lim n^{-1}\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0}$$

$$p\lim n^{-1}\boldsymbol{\varepsilon}'\{\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}'\}\boldsymbol{\varepsilon} = \sigma^2.$$

Urmează (White, Gunst (1979)) că forma limită a lui $\mathbf{A}'\mathbf{A}$ (pentru modelul (1)) este

$$p\lim \mathbf{A}'\mathbf{A} = \Omega = \begin{pmatrix} 1 & \rho' \\ \rho & R \end{pmatrix} \tag{22}$$

unde

$$\rho = R\boldsymbol{\beta}/\{\boldsymbol{\beta}'R\boldsymbol{\beta} + \sigma^2\}^{1/2} \quad \text{şi} \quad R = \mathbf{X}'\mathbf{X}.$$

Deoarece $R = \mathbf{X}'\mathbf{X}$ pot fi cercetate multicolinearităţile în $\mathbf{X}$ în raport de influenţa lor asupra rădăcinilor şi vectorilor proprii ai lui Ω şi deci asupra inferenţelor obţinute din rădăcinile şi vectorii proprii ai lui $\mathbf{A}'\mathbf{A}$ pentru selecţii de volum mare.

Performanţa oricărei analize de regresie va fi examinată cînd $\sigma^2 \to 0$ deoarece în acest caz toate tehnicile vor da predictori perfecţi cînd modelul devine determinist.

Fie $\theta_0 \leqslant \theta_1 \leqslant \ldots \leqslant \theta_k$, $\tau_0, \tau_1, \ldots, \tau_k$, cu $\tau_j' = (\tau_{0j}, \Phi_j')$, $j = 0, 1, \ldots, k$ rădăcinile, respectiv vectorii proprii a lui Ω. Dacă $\sigma^2 = 0$, cea mai mică rădăcină proprie a lui Ω este $\theta_0 = 0$ şi vectorul propriu corespunzător este

$$\tau_0 = \left\{\frac{\boldsymbol{\beta}'R\boldsymbol{\beta}}{\boldsymbol{\beta}'R\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{\beta}}\right\}^{1/2}\begin{pmatrix} 1 \\ -\{\boldsymbol{\beta}'R\boldsymbol{\beta}\}^{-1/2}\boldsymbol{\beta} \end{pmatrix}. \tag{23}$$

Scanned with OKEN Scanner

Pentru a verifica că τ_0 este într-adevăr vectorul propriu al lui Ω ce corespunde la rădăcina proprie $\theta_0 = 0$, notăm că

$$\Omega\tau_0 = \theta_0\tau_0 = 0,$$

ca pentru o rădăcină proprie zero. De asemenea notăm că deoarece R este pozitiv definită, aceasta este singura rădăcină proprie nulă a lui Ω.

Deoarece $\theta_0 = 0$ şi $\sigma_{00} \neq 0$, predictorul perfect pentru $\mathbf{Y}$ este

$$\hat{\mathbf{Y}} = \bar{Y}\mathbf{1} - \eta\mathbf{X}(\tau_{00}^{-1}\Phi_0) =$$

$$= \beta_0\mathbf{1} - \{\boldsymbol{\beta}' R\boldsymbol{\beta}\}^{1/2}\mathbf{X}(-\{\boldsymbol{\beta}' R\boldsymbol{\beta}\}^{-1/2}\boldsymbol{\beta}) = \beta_0\mathbf{1} + \mathbf{X}\boldsymbol{\beta}.$$

În plus, acelaşi rezultat are loc pentru selecţii finite folosind λ_0 şi $\boldsymbol{\gamma}_0$ ai lui $\mathbf{A}'\mathbf{A}$; adică dacă $\sigma^2 = 0$, $\lambda_0 = 0$ însă $\gamma_{00} \neq 0$ găsim un predictor perfect pentru $\mathbf{Y}$ din $\boldsymbol{\gamma}_0$. Implicaţia practică a acestei proprietăţi este importantă pentru analiza regresiei prin rădăcini proprii. Presupunem (pentru acurateţea algoritmului de calcul) în cazul selecţiei de volum finit că λ_0 este zero. Atunci $\gamma_{00} \neq 0$ sau va fi o singularitate în $\mathbf{X}$, contrar presupunerii că $\mathbf{X}$ este de rang maxim coloană. Astfel dacă $\lambda_0 = 0$ obţinem pentru $\mathbf{Y}$ predictorul perfect

$$\hat{\mathbf{Y}} = \bar{Y}\mathbf{1} - \eta\mathbf{X}(\gamma_{00}^{-1}\boldsymbol{\delta}_0).$$

Această situaţie se întîmplă rar în practică. O situaţie mai realistă este aceea în care apare o valoare λ_0 avînd ordinul de mărime mai mic decît al lui l_1, cea mai mică rădăcină proprie a lui $\mathbf{X}'\mathbf{X}$. Dacă λ_0 este de ordin de mărime mai mic decît l_1, atunci $\gamma_{00} \neq 0$. Dacă γ_{00} este zero, vor avea loc relaţiile:

$$\lambda_0 = \boldsymbol{\gamma}_0'\mathbf{A}'\mathbf{A}\boldsymbol{\gamma}_0 = \boldsymbol{\delta}'\mathbf{X}'\mathbf{X}\boldsymbol{\delta}_0 < \mathbf{V}_1'\mathbf{X}'\mathbf{X}\mathbf{V}_1 = l_1.$$

Însă aceasta este o contradicţie deoarece $\mathbf{V}_1$ este singurul vector unitate ce minimizează forma pătratică $\mathbf{V}'\mathbf{X}'\mathbf{X}\mathbf{V}$. Astfel deoarece $\lambda_0 \ll l_1$, γ_{00} nu este zero şi această multicoliniaritate în $\mathbf{A}$ este importantă pentru predicarea lui $\mathbf{Y}$; adică $\boldsymbol{\delta}_0$ conţine informaţii folositoare asupra lui $\boldsymbol{\beta}$ şi va fi reţinut în estimatorul obţinut prin rădăcini proprii.

Dacă $\boldsymbol{\beta} = \mathbf{V}_1$, $\boldsymbol{\delta}_0$ trebuie să conţină informaţii asupra lui $\boldsymbol{\beta}$ dacă $\lambda_0 \leqslant l_1$. Reciproca nu este necesar adevărată, adică dacă $\lambda_0 \simeq l_1$ şi $\boldsymbol{\delta}_0 \simeq \mathbf{V}_1$ atunci $\boldsymbol{\delta}_0$ nu este predictiv şi poate fi înlăturat din estimatorul pentru $\boldsymbol{\beta}$ obţinut prin rădăcini proprii.

Cînd $\boldsymbol{\beta} = c\mathbf{V}_i$, unde $\mathbf{V}_i$ este vectorul propriu corespunzător rădăcinii proprie i a lui $\mathbf{X}'\mathbf{X}$, deoarece $R\mathbf{V}_i = l_i\mathbf{V}_i$,

$$\Omega = \begin{pmatrix} 1 & (c^2 l_i + \sigma^2)^{-1/2} c l_i \mathbf{V}_i \\ (c^2 l_i + \sigma^2)^{-1/2} c l_i \mathbf{V}_i & R \end{pmatrix} \tag{24}$$

Din relaţia $\Omega\tau = \theta\tau$ urmează că cei $k - 1$ vectori proprii ai lui Ω sînt

$$\tau = \begin{pmatrix} 0 \\ \mathbf{V}_p \end{pmatrix} \quad p = 1, 2, \ldots, i-1, i+1, \ldots, k$$

Scanned with OKEN Scanner

corespunzători la rădăcinile proprii $\theta = l_p$. Cei doi vectori proprii rămași sînt de forma :

$$\tau = \begin{pmatrix} a \\ b\mathbf{V}_i \end{pmatrix} \tag{25}$$

ce corespund la rădăcinile proprii :

$$\theta(1) = \{(1 + l_i) - [(1 - l_i)^2 + 4\omega^2]^{1/2}\}/2 \text{ și}$$

$$\theta(2) = \{(1 + l_i) + [(1 - l_i)^2 + 4\omega^2]^{1/2}\}/2, \tag{26}$$

unde $\omega = cl_i(c^2 l_i + \sigma^2)^{-1/2}$. Valorile lui a și b în (25) ce corespund la $\theta(p)$, $p = 1, 2$ sînt

$$a(p) = \omega b(p)/\{\theta(p) - 1\} \text{ și}$$

$$b(p) = \{1 + \omega^2[\theta(p) - 1]^{-2}\}^{-1/2}. \tag{27}$$

Acum presupunem că $\boldsymbol{\beta}'\boldsymbol{\beta}/\sigma^2 = c^2/\sigma^2$ este foarte mare, astfel că în efect $\sigma^2 = 0$ în (24). Atunci rădăcinile proprii (neordonate) ale lui Ω sînt $\theta(1) = 0$, $\theta(2) = 1 + l_i$ și $0 = l_p (p \neq 1)$. Vectorii proprii corespunzător devin

$$(1 + l_i)^{-1/2} \begin{pmatrix} -l_i^{1/2} \\ \mathbf{V}_i \end{pmatrix}, \quad (1 + l_i)^{-1/2} \begin{pmatrix} 1 \\ l_i^{1/2}\mathbf{V}_i \end{pmatrix} \text{ și } \begin{pmatrix} 0 \\ \mathbf{V}_p \end{pmatrix} \text{ pentru } p \neq i. \tag{28}$$

În particular dacă $\boldsymbol{\beta} = \mathbf{V}_k$ rădăcinile proprii ordonate ale lui Ω sînt $0, l_1, l_2, \ldots, l_{k-1}, 1 + l_k$ cu vectorii proprii corespunzători

$$(1 + l_k)^{-1/2} \begin{pmatrix} -l_k^{1/2} \\ \mathbf{V}_k \end{pmatrix}, \begin{pmatrix} 0 \\ \mathbf{V}_1 \end{pmatrix}, \begin{pmatrix} 0 \\ \mathbf{V}_2 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ \mathbf{V}_{k-1} \end{pmatrix},$$

$$\text{și} \quad (1 + l_k)^{-1/2} \begin{pmatrix} 1 \\ l_k^{1/2}\mathbf{V}_k \end{pmatrix}. \tag{29}$$

În (29) sînt importante două aspecte : τ_0 nu conține pe $\mathbf{V}_1$, vectorul propriu ce identifică cea mai puternică multicoliniaritate în $\mathbf{X}'\mathbf{X}$ și $|\tau_{00}|$ nu este mic.

Scanned with OKEN Scanner

Capitolul XXXIV

REGRESIA RIDGE. ESTIMATORI PRIN INVERSĂ GENERALIZATĂ

§ 1 Regresia ridge

1.1. Introducere

Considerăm modelul standard pentru regresia multiplă liniară

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

unde — $\mathbf{X}$ este o matrice $n \times p$ ce conţine valorile a p variabile predictor în fiecare din cele n puncte date şi este de rang p.

— $\boldsymbol{\beta}$ este un vector $p \times 1$, vectorul parametrilor necunoscuţi

— $\boldsymbol{\varepsilon}$ este vectorul erorilor experimentale $n \times 1$, avînd proprietăţile

$$M(\boldsymbol{\varepsilon}) = \mathbf{0} \text{ şi } M(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}_n \tag{2}$$

— $\mathbf{Y}$ este vectorul valorilor observate : $n \times 1$.

M.C.M.M.P. conduce pentru $\boldsymbol{\beta}$ la estimatorul

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{3}$$

cu proprietăţile

$$M(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \text{ şi } \mathscr{C}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{4}$$

Fie L_1 distanţa de la $\hat{\boldsymbol{\beta}}$ la $\boldsymbol{\beta}$. Atunci

$$L_1^2 = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \tag{5}$$

şi

$$M(L_1^2) = \sigma^2 \mathrm{Ur}(\mathbf{X}'\mathbf{X})^{-1}, \tag{6}$$

sau echivalent

$$M(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}'\boldsymbol{\beta} + \sigma^2 \mathrm{Ur}(\mathbf{X}'\mathbf{X})^{-1}. \tag{6'}$$

Scanned with OKEN Scanner

Cînd ε este repartizat normal

$$D^2(L_1^2) = 2\sigma^4 \mathrm{Ur}(\mathbf{X}'\mathbf{X})^{-2}, \tag{7}$$

ceea ce arată că $\hat{\boldsymbol{\beta}}$ este nesigur cînd $\mathbf{X}'\mathbf{X}$ este depărtată de matricea unitate.

Dacă valorile proprii ale matricii $\mathbf{X}'\mathbf{X}$ sînt notate prin

$$\lambda_{max} = \lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p = \lambda_{min} > 0 \tag{8}$$

atunci valoarea medie a pătratului distanței dintre $\hat{\boldsymbol{\beta}}$ și $\boldsymbol{\beta}$ este dată de

$$M(L_1^2) = \sigma^2 \sum_{i=1}^{p} (1/\lambda_i) \tag{9}$$

și dispersia, cînd eroarea este normală, de

$$D^2(L_1^2) = 2\sigma^4 \sum_{i=1}^{p} (1/\lambda_i^2). \tag{10}$$

Urmează că marginile inferioare pentru media și dispersia pătratului distanței de la $\hat{\boldsymbol{\beta}}$ la $\boldsymbol{\beta}$ sînt

$$\sigma^2/\lambda_{min} \text{ și respectiv } 2\sigma^4/\lambda_{min}. \tag{11}$$

Deci dacă $\mathbf{X}'\mathbf{X}$ are una sau mai multe valori proprii mici, distanța dintre $\hat{\boldsymbol{\beta}}$ și $\boldsymbol{\beta}$ va tinde să se mărească.

1.2. Regresia ridge

DEFINIȚIA 1. (Hoerl (1962), Hoerl, Kennard (1970 a)). Statistica

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y} \tag{12}$$

pentru $k \geqslant 0$ poartă numele de *estimația ridge* a parametrului $\boldsymbol{\beta}$ și este soluția ecuației

$$(\mathbf{X}'\mathbf{X} + k\mathbf{I})\hat{\boldsymbol{\beta}}^* = \mathbf{X}'\mathbf{Y}. \tag{13}$$

Punînd

$$(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} = \mathbf{W}, \tag{14}$$

(12) se scrie

$$\hat{\boldsymbol{\beta}}^* = \mathbf{W}\mathbf{X}'\mathbf{Y}. \tag{15}$$

PROPOZIȚIA 1. (Hoerl, Kennard (1970 a)). *Legătura dintre estimația ridge și estimația de C.M.M.P. este dată de*

$$\hat{\boldsymbol{\beta}}^* = [\mathbf{I}_p + k(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\hat{\boldsymbol{\beta}} \tag{16}$$

Scanned with OKEN Scanner

sau punînd

$$[\mathbf{I}_p + k(\mathbf{X}'\mathbf{X})^{-1}]^{-1} = \mathbf{Z} \tag{17}$$

de

$$\hat{\boldsymbol{\beta}}^* = \mathbf{Z}\hat{\boldsymbol{\beta}}. \tag{18}$$

Observație. Cu alte cuvinte estimația $\hat{\boldsymbol{\beta}}^*$ este o transformare liniară a estimației $\hat{\boldsymbol{\beta}}$ și transformarea depinde numai de $\mathbf{X}$ și k.

Demonstrația rezultă din (12) și din faptul că $\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}}$. Proprietățile statisticilor $\mathbf{W}$ și $\mathbf{Z}$ care vor fi utilizate sînt date în

PROPOZIȚIA 2. *Dacă* $\xi_i(\mathbf{W})$ *și* $\xi_i(\mathbf{Z})$ *sînt valorile proprii ale lui* $\mathbf{W}$ *și respectiv* $\mathbf{Z}$, *atunci*

$$\xi_i(\mathbf{W}) = 1/(\lambda_i + k), \tag{19}$$

$$\xi_i(\mathbf{Z}) = \lambda_i/(\lambda_i + k), \tag{20}$$

unde λ_i *sînt valorile proprii ale matricii* $\mathbf{X}'\mathbf{X}$.

Demonstrația rezultă din (14), (17), (15) și (18) și din soluțiile ecuațiilor caracteristice

$$|\mathbf{W} - \xi\mathbf{I}| = \mathbf{0} \quad \text{și} \quad |\mathbf{Z} - \xi\mathbf{I}| = \mathbf{0}.$$

PROPOZIȚIA 3. *Avem:*

$$\mathbf{Z} = [\mathbf{I}_p + k(\mathbf{X}'\mathbf{X})^{-1}]^{-1} = \mathbf{I} - k\mathbf{W}. \tag{21}$$

Demonstrație. Putem scrie

$$\mathbf{Z} = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}$$

sau ținînd seama de (14)

$$\mathbf{Z} = \mathbf{W}\mathbf{X}'\mathbf{X}.$$

Cum

$$\mathbf{W}\mathbf{X}'\mathbf{X} = \mathbf{I} - k\mathbf{W},$$

propoziția este demonstrată.

PROPOZIȚIA 4 (Hoerl, Kennard (1970 a)). $\hat{\boldsymbol{\beta}}^*$, *pentru* $k \neq 0$, *este mai scurtă decît* $\hat{\boldsymbol{\beta}}$, *adică*

$$(\hat{\boldsymbol{\beta}}^*)'\hat{\boldsymbol{\beta}}^* < \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}. \tag{22}$$

Demonstrație. Din (18)

$$\hat{\boldsymbol{\beta}}^* = \mathbf{Z}\hat{\boldsymbol{\beta}}$$

și din presupunerile asupra lui $\mathbf{X}'\mathbf{X}$, $\mathbf{Z}$ este simetrică și pozitiv definită. Atunci

$$(\hat{\boldsymbol{\beta}}^*)'\hat{\boldsymbol{\beta}}^* \leqslant \xi^2_{\max}(\mathbf{Z})\,\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} \text{ (Riley (1955)).} \tag{23}$$

Scanned with OKEN Scanner

Însă $\xi_{max}(\mathbf{Z}) = \lambda_1/(\lambda_1 + k)$ unde λ_1 este cea mai mare valoare proprie a matricii $\mathbf{X}'\mathbf{X}$ și (22) este demonstrată.

Observație. Din (20) și (21),

$$\mathbf{Z}(0) = \mathbf{I}$$

și $\mathbf{Z}$ se apropie de $\mathbf{0}$ cînd $k \to \infty$.

PROPOZIȚIA 5. (Hoerl, Kennard (1970 a), Marquardt (1970)). *Fie $k \geqslant 0$ arbitrar și $\hat{\boldsymbol{\beta}}^*$ satisfăcînd* (12). *Atunci $\hat{\boldsymbol{\beta}}^*$ minimizează* SS *reziduală*

$$\phi(\hat{\boldsymbol{\beta}}^*) = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*)'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*) \tag{24}$$

pe sfera centrată în origine avînd raza la pătrat egală cu $(\hat{\boldsymbol{\beta}}^*)'\,\hat{\boldsymbol{\beta}}^*$.

Observație. Cum (24) se mai scrie

$$\phi(\hat{\boldsymbol{\beta}}^*) = \mathbf{Y}'\mathbf{Y} - (\hat{\boldsymbol{\beta}}^*)'\mathbf{X}'\mathbf{Y} - k(\hat{\boldsymbol{\beta}}^*)'\hat{\boldsymbol{\beta}}^*$$

urmează că $\phi(\hat{\boldsymbol{\beta}}^*)$ este o funcție monoton descrescătoare de k. Astfel, soluția ridge cere o creștere a SS reziduală față de SS reziduală a C.M.M.P.

PROPOZIȚIA 6. (Marquardt (1970)). *Fie $\hat{\boldsymbol{\beta}}^*$ soluția lui* (13) *pentru o valoare dată a lui k. Atunci $\| \hat{\boldsymbol{\beta}}^* \|$ este o funcție monoton descrescătoare de k, astfel încît pentru $k \to \infty$, $\| \hat{\boldsymbol{\beta}}^* \| \to 0$.*

PROPOZIȚIA 7. (Marquardt (1963), (1970)). *In* (13), $\mathbf{g} = \mathbf{X}'\mathbf{Y}$ *este vectorul gradient al lui $\phi(\hat{\boldsymbol{\beta}})$. Fie γ_k unghiul dintre $\hat{\boldsymbol{\beta}}^*$ și $\mathbf{g}$. Atunci γ_k este o funcție continuă, monoton descrescătoare de k, astfel încît cînd $k \to \infty$, $\gamma_k \to 0$. Deoarece $\mathbf{g}$ este independent de k, urmează că $\hat{\boldsymbol{\beta}}^*$ se rotește în direcția lui $\mathbf{g}$ cînd $k \to \infty$.*

1.3. Urma ridge

1.3.1. *Definiția urmei ridge. Caracterizarea urmei ridge.* Cînd $\mathbf{X}'\mathbf{X}$ deviază considerabil de la matricea unitate, adică cînd are valori proprii mici, (6) și (7) arată că probabilitatea ca $\hat{\boldsymbol{\beta}}$ să fie apropiată de $\boldsymbol{\beta}$ poate să fie mică.

Calculînd $\hat{\boldsymbol{\beta}}^*(k)$ și $\phi^*(k)$ pentru o mulțime de valori ale lui k se pot obține rezultate privind structura spațiului factori și sensibilitatea soluției pentru mulțimi particulare de date.

Fie $\mathbf{B}$ o estimație oarecare a vectorului $\boldsymbol{\beta}$. Atunci SS reziduală poate fi scrisă ca

$$\begin{aligned} \phi &= (\mathbf{Y} - \mathbf{X}\mathbf{B})'(\mathbf{Y} - \mathbf{X}\mathbf{B}) = \\ &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\mathbf{B} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\mathbf{B} - \hat{\boldsymbol{\beta}}) = \\ &= \phi_{min} + \phi(\mathbf{B}). \end{aligned} \tag{25}$$

Scanned with OKEN Scanner

Conturele de ϕ constant sînt suprafeţele hiperelipsoizilor centraţi în $\hat{\boldsymbol{\beta}}$, estimatorul obişnuit de C.M.M.P. pentru $\boldsymbol{\beta}$.

Valoarea lui ϕ este valoarea minimă $\phi_{\min}$, plus valoarea formei pătratice în $(\mathbf{B} - \boldsymbol{\beta})$. Există un interval de valori ale lui $\mathbf{B}_0$ care satisface relaţia

$$\phi = \phi_{\min} + \phi_0,$$

unde $\phi_0 > 0$ este o creştere fixată.

Urma ridge poate fi descrisă ca un drum pe suprafaţa SS astfel că pentru un ϕ fixat este aleasă o singură valoare a lui $\mathbf{B}$ şi aceasta este una cu lungima minimă.

Aceasta se poate stabili precis după cum urmează:

Minimizăm $\mathbf{B}'\mathbf{B}$ cu condiţia

$$(\mathbf{B} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\mathbf{B} - \hat{\boldsymbol{\beta}}) = \phi_0. \tag{26}$$

Ca problemă Lagrangian aceasta este

$$\text{Minimizăm } F = \mathbf{B}'\mathbf{B} + (1/k)[(\mathbf{B} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\mathbf{B} - \hat{\boldsymbol{\beta}})' - \phi_0] \tag{27}$$

unde $(1/k)$ este multiplicatorul. Atunci

$$\frac{\partial F}{\partial \mathbf{B}} = 2\mathbf{B} + (1/k)[2(\mathbf{X}'\mathbf{X})\,\mathbf{B} - 2(\mathbf{X}'\mathbf{X})\,\hat{\boldsymbol{\beta}}] = 0. \tag{28}$$

Aceasta se reduce la

$$\mathbf{B} = \hat{\boldsymbol{\beta}}^* = [\mathbf{X}'\mathbf{X} + k\mathbf{I}]^{-1}\mathbf{X}'\mathbf{Y}, \tag{29}$$

unde k este ales astfel încît să satisfacă (26). Acesta este estimatorul ridge. În practică se poate alege $k \geqslant 0$ şi se calculează ϕ_0.

În termeni de $\hat{\boldsymbol{\beta}}^*$, SS reziduală devine

$$\phi^*(k) = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*)'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*) = \phi_{\min} + k^2\hat{\boldsymbol{\beta}}^{*\prime}(\mathbf{X}'\mathbf{X})^{-1}\hat{\boldsymbol{\beta}}^* \tag{30}$$

O presupunere complet echivalentă este:

Dacă pătratul lungimii vectorului regresiei $\mathbf{B}$ este fixată (R^2), atunci $\hat{\boldsymbol{\beta}}^*$ este valoarea lui $\mathbf{B}$ care realizează minimul SS. Adică $\hat{\boldsymbol{\beta}}^*$ este valoarea lui $\mathbf{B}$ care minimizează funcţia

$$F_1 = (\mathbf{Y} - \mathbf{X}\mathbf{B})'(\mathbf{Y} - \mathbf{X}\mathbf{B}) + (1/k)(\mathbf{B}'\mathbf{B} - R^2). \tag{31}$$

1.3.2. *Caracterizarea prin verosimilităţi a urmei ridge.* Folosind ipoteza conform căreia vectorul erorilor este repartizat $N(0, \sigma^2\mathbf{I}_n)$ funcţia de verosimilitate este

$$(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right\}. \tag{32}$$

Scanned with OKEN Scanner

Nucleul acestei funcţii este forma pătratică din exponenţială care poate fi scrisă sub forma

$$(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}). \tag{33}$$

Din (25) aceasta arată că o creştere în SS reziduală este echivalentă cu o descreştere în valoare a funcţiei de verosimilitate. Urma ridge poate fi interpretată ca un drum în spaţiul verosimilităţilor şi se pune problema dacă acest drum particular poate fi de un interes. Motivaţia este aceeaşi ca pentru SS. Cu toate că vectorii lungi dau aceleaşi valori ale verosimilităţii ca şi vectorii scurţi, ei nu vor avea întotdeana aceeaşi semnificaţie fizică. Implicit este o restrîngere a valorilor posibile ale lui $\hat{\boldsymbol{\beta}}$ care nu este făcută explicit în formularea modelului liniar general (1). Această implicaţie este discutată în cele ce urmează.

1.4. Proprietăţi ale mediei erorii pătratice a regresiei ridge

Pentru a privi $\hat{\boldsymbol{\beta}}^*$ din punctul de vedere al medii erorii pătratice este necesar să obţinem o expresie pentru $M(L_1^2(k))$. Aplicarea operatorului medie şi (16) dau

$$M[L_1^2(k)] = M[(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})] = \tag{34}$$

$$= M[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{Z}'\mathbf{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] + (\mathbf{Z}\boldsymbol{\beta} - \boldsymbol{\beta})'(\mathbf{Z}\boldsymbol{\beta} - \boldsymbol{\beta})$$

sau

$$M[L_1^2(k)] = \sigma^2 \mathrm{Ur}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Z} + \boldsymbol{\beta}'(\mathbf{Z} - \mathbf{I})'(\mathbf{Z} - \mathbf{I})\boldsymbol{\beta}. \tag{35}$$

În continuare

$$M[L_1^2(k)] = \sigma^2[\mathrm{Ur}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} - k\mathrm{Ur}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-2}] +$$

$$+ k^2\boldsymbol{\beta}'(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-2}\boldsymbol{\beta},$$

şi

$$M[L_1^2(k)] = \sigma^2 \sum_1^p \lambda_i/(\lambda_i + k)^2 + k^2\boldsymbol{\beta}'(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-2}\boldsymbol{\beta} = \tag{37}$$

$$= \gamma_1(k) + \gamma_2(k). \tag{38}$$

$\gamma_2(k)$ este pătratul distanţei de la $\mathbf{Z}\boldsymbol{\beta}$ la $\boldsymbol{\beta}$. El va fi zero cînd $k = 0$, deoarece în acest caz $\mathbf{Z} = \mathbf{I}$. Astfel, $\gamma_2(k)$ poate fi considerat pătratul deplasării introduse prin folosirea lui $\hat{\boldsymbol{\beta}}^*$ în locul lui $\hat{\boldsymbol{\beta}}$.

$\gamma_1(k)$ este suma dispersiilor (dispersia totală) parametrilor estimaţi.

În termeni de variabila aleatoare $\mathbf{Y}$

$$\hat{\boldsymbol{\beta}}^* = \mathbf{Z}\hat{\boldsymbol{\beta}} = \mathbf{Z}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{39}$$

şi are loc

PROPOZIŢIA 8. *Matricea de covariante a estimatorului* $\hat{\boldsymbol{\beta}}^*$ *este*

$$\mathscr{C}(\hat{\boldsymbol{\beta}}^*) = \mathbf{Z}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathscr{C}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{Z}' = \tag{40}$$

$$= \sigma^2\mathbf{Z}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{Z}'$$

Scanned with OKEN Scanner

Observație. Suma dispersiilor tuturor statisticilor $\hat{\beta}_i^*$ este suma elementelor diagonale ale lui (40).

TEOREMA 9 (Hoerl, Kennard (1970 a)). *Dispersia totală* $\gamma_1(k)$ *este o funcție continuă, monoton descrescătoare de* k.

COROLARUL 1. *Prima derivată în raport cu* k *a dispersiei totale* : $\gamma_1'(k)$ *tinde către* $-\infty$ *cînd* $k \to 0^+$ *și* $\lambda_p \to 0$.

TEOREMA 10 (Hoerl, Kennard (1970 a)). *Pătratul deplasării* $\gamma_2(k)$ *este o funcție continuă și monoton descrescătoare de* k.

Demonstrație. Din (37)

$$\gamma_2(k) = k^2\boldsymbol{\beta}'(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-2}\boldsymbol{\beta}.$$

Dacă Λ este matricea valorilor proprii ale lui $\mathbf{X}'\mathbf{X}$ și $\mathbf{P}$ este transformarea ortogonală astfel încît

$$\mathbf{X}'\mathbf{X} = \mathbf{P}'\Lambda\mathbf{P},$$

atunci

$$\gamma_2(k) = k^2 \sum_1^p \alpha_i^2/(\lambda_i + k)^2, \tag{41}$$

unde

$$\boldsymbol{\alpha} = \mathbf{P}\boldsymbol{\beta}. \tag{42}$$

Întrucît $\lambda_i > 0$ pentru orice i și $k \geqslant 0$, fiecare element $\lambda_i + k$ este pozitiv și nu există singularități în sumă. $\gamma_2(0) = 0$. Atunci $\gamma_2(k)$ este o funcție continuă pentru $k \geqslant 0$. Pentru $k > 0$, (41) poate fi scrisă ca

$$\gamma_2(k) = \sum_1^p \alpha_i^2/[1 + (\lambda_i/k)]^2. \tag{43}$$

Întrucît $\lambda_i > 0$ pentru toți i, funcțiile λ_i/k sînt monoton descrescătoare pentru k crescător și fiecare termen al lui $\gamma_2(k)$ este monoton crescător. Astfel $\gamma_2(k)$ este monoton crescătoare.

COROLARUL 2. *Pătratul deplasării* $\gamma_2(k)$ *aproximează* $\boldsymbol{\beta}'\boldsymbol{\beta}$ *ca o limită superioară.*

Demonstrație. Din (41)

$$\lim_{k\to\infty} \gamma_2(k) = \sum_1^p \alpha_i^2 = \boldsymbol{\alpha}'\boldsymbol{\alpha} = \boldsymbol{\beta}'\mathbf{P}'\mathbf{P}\boldsymbol{\beta} = \boldsymbol{\beta}'\boldsymbol{\beta}.$$

COROLARUL 3. *Derivata* $\gamma_2'(k)$ *tinde la zero cînd* $k \to 0^+$.

Demonstrație. Din (41) se stabilește că

$$\frac{d\gamma_2(k)}{dk} = 2k \sum_1^p \lambda_i\alpha_i^2/(\lambda_i + k)^3. \tag{44}$$

Fiecare termen al sumei $2k\lambda_i\alpha_i^2/(\lambda_i + k)^3$ este o funcție continuă și este zero cînd $k \to 0^+$.

TEOREMA 11. (Hoerl, Kennard (1970 a)) (Teorema de existenţă). *Există întotdeauna un* $k>0$ *astfel încît*

$$M[L_1^2(k)]<M[L_1^2(0)] = \sigma^2 \sum_1^p (1/\lambda_i).$$

Demonstraţie. Din (37), (41), şi (44)

(45) $$\mathrm{d}M[L_1^2(k)]/\mathrm{d}k = \mathrm{d}\gamma_1(k)/\mathrm{d}k + \mathrm{d}\gamma_2(k)/\mathrm{d}k =$$
$$= -2\sigma^2 \sum_1^p \lambda_i/(\lambda_i+k)^3 + 2k \sum_1^p \lambda_i\alpha_i^2/(\lambda_i+k)^3.$$

Mai întîi observăm că

$$\gamma_1(0) = \sigma^2 \sum_1^p (1/\lambda_i) \quad \text{şi} \quad \gamma_2(0) = 0.$$

În teoremele 9 şi 10 s-a stabilit că $\gamma_1(k)$ şi $\gamma_2(k)$ sînt monoton descrescătoare şi respectiv crescătoare. Primele lor derivate sînt întotdeauna nepozitive şi respetiv nenegative. Astfel, pentru a demonstra teorema este necesar numai să arătăm că există întotdeauna un $k>0$ astfel încît

$$\mathrm{d}M[L_1^2(k)]/\mathrm{d}k<0.$$

Condiţia pentru aceasta este arătată de (45) ca fiind

(46) $$k<\sigma^2/\alpha_{\max}^2.$$

1.5. Forma generală a regresiei ridge

Este întotdeauna posibil să reducem problema generală de regresie liniară definită în (1) la o formă canonică în care matricea $\mathbf{X'X}$ este diagonală. În particular, există o transformare ortogonală $\mathbf{P}$ astfel încît

(47) $$\mathbf{X'X} = \mathbf{P'} \wedge \mathbf{P},$$

unde $\Lambda = (\delta_{ij}\lambda_i)$ este matricea valorilor proprii ale lui $\mathbf{X'X}$. Fie

(48) $$\mathbf{X} = \mathbf{X^*P}$$

şi

(49) $$\mathbf{Y} = \mathbf{X^*\alpha} + \boldsymbol{\varepsilon}$$

unde

(50) $$\boldsymbol{\alpha} = \mathbf{P\beta}, \ (\mathbf{X^*})'\mathbf{X^*} = \Lambda \text{ şi } \boldsymbol{\alpha}'\boldsymbol{\alpha} = \boldsymbol{\beta}'\boldsymbol{\beta}.$$

Atunci, procedeul general de estimaţie ridge este definit de

(51) $$\boldsymbol{\alpha}^* = [(\mathbf{X^*})'(\mathbf{X^*}) + \mathbf{K}]^{-1}(\mathbf{X^*})'\mathbf{Y},$$

unde

$$\mathbf{K} = (\delta_{ij}k_i), \ k_i > 0.$$

Scanned with OKEN Scanner

Toate rezultatele de bază formulate anterior rămîn adevărate și pentru acest caz mai general.

Mai important este faptul că are loc un echivalent al teoremei de existență.

În forma generală se caută un k_i pentru fiecare variabilă canonică definită prin $\mathbf{X}^*$. Definind

$$(L_1^*)^2 = (\hat{\boldsymbol{\alpha}}^* - \boldsymbol{\alpha})'(\hat{\boldsymbol{\alpha}}^* - \boldsymbol{\alpha})$$

se arată că valorile optime pentru k_i sînt σ^2/α_i^2.

Bazat pe experiență, cea mai bună metodă pentru a obține o estimație superioară $\hat{\boldsymbol{\beta}}^*$ este folosirea unui $k_i = k$ pentru toți i și a urmei ridge pentru a selecta o singură valoare k și un $\hat{\boldsymbol{\beta}}^*$ unic.

Regresia ridge a apărut din contactul cu alte abordări ale analizei regresiei și cu lucrări cu același scop. Vom menționa două dintre acestea.

Într-o serie de lucrări Stein ((1960), (1962)) și James, Stein (1961) cercetează modificarea în media erorii pătratice printr-o transformare asupra lui $\hat{\boldsymbol{\beta}}$ de forma

$$C\hat{\boldsymbol{\beta}}, \quad 0 \leqslant C \leqslant 1,$$

care este o scurtare a vectorului $\hat{\boldsymbol{\beta}}$. Ei arată că un astfel de $C > 0$ poate fi găsit întotdeauna și indică cum poate fi calculat.

Balakrishnan (1963) consideră estimația cu constrîngeri în legătură cu regresia. Presupune că $\boldsymbol{\beta}$ este constrîns să aparțină mulțimii convexe mărginită și închisă $\mathbf{C}$ (în particular hipersfera de rază R). Consideră criteriu de estimare: minimul SS reziduale

$$\phi = (\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB}),$$

unde $\mathbf{B}$ este valoarea ce dă minimul. În constrîngeri,

$$\text{dacă } \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} \leqslant R^2 \text{ atunci } \mathbf{B} \text{ este ales să fie } \hat{\boldsymbol{\beta}};$$

în rest $\mathbf{B}$ este ales să fie $\hat{\boldsymbol{\beta}}^*$ unde k este ales astfel încît

$$(\hat{\boldsymbol{\beta}}^*)'(\hat{\boldsymbol{\beta}}^*) = R^2.$$

Regresia ridge are două aspecte

Primul este urma ridge care este o reprezentare în două dimensiuni a lui $\hat{\boldsymbol{\beta}}_i^*(k)$ și SS reziduală $\phi^*(k)$ pentru un număr de valori ale lui k în intervalul [0, 1]. Urma servește pentru a descrie interrelațiile complexe ce există între vectori de predicție neortogonali și efectele acestor interrelații asupra estimației lui $\boldsymbol{\beta}$.

Al doilea aspect este determinat de valorile lui k care dau o estimație mai bună pentru $\boldsymbol{\beta}$.

§ 2. Estimatori ai regresiei prin inversă generalizată

2.1 Inverse generalizate

În problemele de estimare apare matricea $\mathbf{A} = \mathbf{X}'\mathbf{X}$ ale cărei valori proprii pot fi grupate calitativ în trei tipuri: substanțial mai mari decît zero, ușor mai mari decît zero, precis zero (exceptînd erorile de rotunjire).

Scanned with OKEN Scanner

În calcule este adeseori dificil să distingem între tipurile adiacente. Obişnuit se observă un spectru de valori proprii ce variază de la valori mari şi pozitive la foarte mici sau zero.

Cum apare posibilitatea singularităţii lui **A**, încercăm să inversăm **A** cu ajutorul inversei generalizate (Greville (1959), (1960), Rao (1962), (1965)).

Existînd mai multe metode pentru determinarea inversei generalizate, urmînd Marquardt (1970) vom proceda după cum urmează.

Fie **D** matricea diagonalizată, cu elementele diagonale ordonate

$$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p,$$

şi **S** matricea vectorului propriu ce transformă **A** în **D**. Astfel

$$\mathbf{S'AS} = \mathbf{D} \tag{52}$$

unde $\mathbf{S'S} = \mathbf{I}$. Atunci

$$\mathbf{A^{-1}} = \mathbf{SD^{-1}S'}. \tag{53}$$

Presupunem că **A** este de rang r astfel că ultimele $(p - r)$ elemente ordonate ale lui **D** sînt zero. Partiţionăm **S** după cum urmează :

$$\mathbf{S} = (\mathbf{S}_r \mid \mathbf{S}_{p-r}) \tag{54}$$

unde $\mathbf{S}_r$ este $(p \times r)$; $\mathbf{S}_{p-r}$ este $[p \times (p - r)]$.
Partiţionăm **D** similar

$$D = \left(\begin{array}{c|c} \mathbf{D}_r & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{D}_{p-r} \end{array}\right) \tag{55}$$

unde $\mathbf{D}_r$ este $r \times r$, D_{p-r} este $(p - r) \times (p - r)$.
Acum, prin presupunere $\mathbf{D}_{p-r}$ este zero astfel ca $\mathbf{D}^{-1}_{p-r} = \mathbf{0}$. Astfel, inversa devine

$$\mathbf{A}_r^+ = \mathbf{S}_r\mathbf{D}_r^{-1}\mathbf{S}_r'. \tag{56}$$

(56) se poate scrie

$$\mathbf{A}_r^+ = \sum_{j=1}^{r} \frac{1}{\lambda_j} \mathbf{S}_j\mathbf{S}_j' \tag{57}$$

unde $\mathbf{S}_j$ este vectorul propriu al lui **A** ce corespunde la λ_j. Calcularea rangului r a lui **A** este bazată pe spectrul de valori proprii λ_j. Reamintind că urma este păstrată prin transformările ortonormale (52) şi că λ_j este suma pătratelor proiecţiilor punctelor reprezentate prin liniile lui **X** în vectorul propriu asociat cu valoarea proprie (valorile fiind ordonate) j, va fi adoptat un criteriu astfel ca rangul atribuit va include substanţial toate variaţiile în punctele lui **X**. Un astfel de criteriu este oferit de cea mai mică valoare a lui r pentru care

$$\left| \frac{\sum_{j=p}^{p-r} \lambda_j}{\mathrm{Ur}\mathbf{D}} \right| < \omega \tag{58}$$

Scanned with OKEN Scanner

unde în diferitele aplicaţii ω poate fi selectat din intervalul 10^{-1} la 10^{-7}, obişnuit 10^{-5}.

Observaţie. Inversa (57) fiind nestabilă datorată prezenţei valorilor proprii λ_j care sînt apropiate de zero, putem creşte rangul lui **A** prin adăugarea unei cantităţi pozitive mici lui λ_j în numitorul fiecărui termen din (57). Acest procedeu este precis estimatorul ridge discutat în paragrafele anterioare.

Marquardt (1970) cercetează o clasă de estimatori prin inversă generalizată cînd rangul atribuit poate fi un întreg $1 \leqslant r \leqslant p$ reprezentînd numărul termenilor păstraţi în (57), sau o variabilă continuă $0 \leqslant r \leqslant p$ obţinută prin includerea tuturor termenilor în (57) pentru care $j \leqslant$ partea întreagă a lui r.

Cazul cînd r este un întreg a fost numit regresia *componentelor principale* sau *regresia ortogonală* (Kendall (1957)).

2.2. Proprietăţi ale soluţiei inversă generalizată

DEFINIŢIA 2 (Marquardt (1970)) Estimatorii

$$\hat{\boldsymbol{\beta}}^{+} = \mathbf{A}_r^{+}\mathbf{X}'\mathbf{Y} \tag{59}$$

unde $\mathbf{A}_r^{+}$ este inversa generalizată (57) poartă numele de *estimatori ai regresiei prin inversa generalizată* (sau estimatori Marquardt).

În general există o valoare *optimă* a lui r pentru orice problemă, însă prezintă importanţă examinarea soluţiei prin inversa generalizată pentru un domeniu de valori admisibile pentru r.

TEOREMA 12 (Marquardt (1970)). *Fie* $\hat{\boldsymbol{\beta}}^{+}$ *soluţia ecuaţiilor normale*

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$$

obţinută prin atribuirea rangului r *matricei* $\mathbf{A} = \mathbf{X}'\mathbf{X}$ *şi folosind inversa generalizată* (57). *Atunci* $\hat{\boldsymbol{\beta}}^{+}$ *minimizează* SS *reziduriilor*

$$\phi(\hat{\boldsymbol{\beta}}) = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

pentru toţi $\hat{\boldsymbol{\beta}}$ *în subspaţiul* r*-dimensional.*

TEOREMA 13. *Fie* $\hat{\boldsymbol{\beta}}^{+}$ *definit ca în teorema* 12. *Atunci* $\|\hat{\boldsymbol{\beta}}^{+}\|^2$ *este o funcţie crescătoare de* r *în trepte.*

TEOREMA 14. (Marquardt (1970)). *Fie* γ_r *unghiul dintre* $\hat{\boldsymbol{\beta}}^{+}$ *şi* $\mathbf{g} = \mathbf{A}\hat{\boldsymbol{\beta}}^{+}$. *Atunci* $\gamma_r \geqslant \gamma_{r-1}$ (r *un întreg*) λ_r *şi* λ_{r-1} *satisfac inegalităţile*

$$(0 < \lambda_r), \quad (\lambda_r/\lambda_{r-1} \leqslant 1), \quad (\lambda_{r-1}/\lambda_{r-2} \leqslant 1).$$

Deoarece $\mathbf{g}$ este independentă de $\hat{\boldsymbol{\beta}}^{+}$, urmează că $\hat{\boldsymbol{\beta}}^{+}$ se roteşte în direcţia lui $\mathbf{g}$ cînd r descreşte.

TEOREMA 15. (Marquardt (1970)). *Estimaţia* $\hat{\boldsymbol{\beta}}^{+}$ *este o tansformare liniară de* $\hat{\boldsymbol{\beta}}$ *şi transformarea depinde numai de* $\mathbf{X}$ *şi* r.

Scanned with OKEN Scanner

Demonstrație. Din (59) și (56)

$$\hat{\boldsymbol{\beta}}^{+} = \mathbf{S}_r\mathbf{D}_r^{-1}\mathbf{S}_r'\mathbf{X}'\mathbf{Y}. \tag{60}$$

Însă

$$\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X})\,\hat{\boldsymbol{\beta}}.$$

Astfel

$$\hat{\boldsymbol{\beta}}^{+} = \mathbf{S}_r\mathbf{D}_r^{-1}\mathbf{S}_r'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{Z}_r\hat{\boldsymbol{\beta}}. \tag{61}$$

Urmează că $\hat{\boldsymbol{\beta}}^{+}$ este un estimator deplasat pentru $\boldsymbol{\beta}$ (dacă $\mathbf{D}_{p-r}$ este o matrice nenulă).

TEOREMA 16. *Matricea de covarianțe a lui* $\hat{\boldsymbol{\beta}}^{+}$ *este*

$$\mathscr{C}(\hat{\boldsymbol{\beta}}^{+}) = \sigma^2[\mathbf{S}_r\mathbf{D}_r^{-1}\mathbf{S}_r'](\mathbf{X}'\mathbf{X})\;[\mathbf{S}_r\mathbf{D}_r^{-1}\mathbf{S}_r']. \tag{62}$$

Demonstrație. Cum

$$\mathscr{C}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1},$$

$$\mathscr{C}(\mathbf{Z}_r\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{Z}_r(\mathbf{X}'\mathbf{X})^{-1}\mathbf{Z}_r',$$

sau ținînd seama de (61) rezultă (62).

TEOREMA 17 (Marquardt (1970)). *Eroarea medie pătratică a lui* $\hat{\boldsymbol{\beta}}^{+}$ *este*

$$M(L_1^2) = \mathrm{Ur}[\mathscr{C}(\hat{\boldsymbol{\beta}}^{+})] + \boldsymbol{\beta}'(\mathbf{Z}_r - \mathbf{I})'(\mathbf{Z}_r - \mathbf{I})\,\boldsymbol{\beta}. \tag{63}$$

Observație. Al doilea termen din (63) este pătratul deplasării. Va fi zero cînd $r = p$.

COROLARUL 4. *Primul termen din* (63) *este o funcție crescătoare de* r.

Demonstrație. Din (57) avem:

$$\mathbf{S}_r\mathbf{D}_r^{-1}\mathbf{S}_r' = \sum_{j=1}^{r}\frac{1}{\lambda_j}\mathbf{S}_j\mathbf{S}_j'. \tag{64}$$

Deci

$$\mathrm{Ur}[\mathbf{S}_r\mathbf{D}_r^{-1}\mathbf{S}_r'] = \sum_{j=1}^{r}\frac{1}{\lambda_j}\mathrm{Ur}(\mathbf{S}_j\mathbf{S}_j').$$

Însă

$$\mathrm{Ur}(\mathbf{S}_j\mathbf{S}_j') = \|\mathbf{S}_j\| = 1,$$

deoarece $\mathbf{S}$ este o rotație ortonormală. Astfel

$$\mathrm{Ur}[\mathbf{S}_r\mathbf{D}_r^{-1}\mathbf{S}_r'] = \sum_{j=1}^{r}\frac{1}{\lambda_j} \tag{65}$$

$$\mathrm{Ur}[\mathscr{C}(\hat{\boldsymbol{\beta}}^{+})] = \sigma^2\left(\sum_{j=1}^{r}\frac{1}{\lambda_j}\right). \tag{66}$$

Scanned with OKEN Scanner

Deoarece $\lambda_j \geq 0$ pentru toţi j, (66) creşte monoton cu r. În cazul special cînd datele sînt ortogonale, adică cînd toţi $\lambda_j = 1$ şi cînd $r = p$, acest rezultat devine $p\sigma^2$.

COROLARUL 5. *Termenul deplasare în $M(L_1^2)$ este o funcţie monoton descrescătoare de r.*

Demonstraţie. Avem

$$(Deplasarea)^2 = \boldsymbol{\beta}'(\mathbf{Z}_r - \mathbf{I})'(\mathbf{Z}_r - \mathbf{I})\boldsymbol{\beta} =$$

$$= \boldsymbol{\beta}'[\mathbf{S}_r\mathbf{D}_r^{-1}\mathbf{S}_r'(\mathbf{X}'\mathbf{X}) - \mathbf{I}]'[\mathbf{S}_r\mathbf{D}_r^{-1}\mathbf{S}_r'(\mathbf{X}'\mathbf{X}) - \mathbf{I}]\boldsymbol{\beta}$$

şi cum

$$(\mathbf{Z}_r - \mathbf{I}) = \mathbf{S}_r\mathbf{S}_r' - \mathbf{I}_p = -\mathbf{S}_{p-r}\mathbf{S}_{p-r}', \tag{67}$$

$$(Deplasarea)^2 = \boldsymbol{\beta}'[\mathbf{S}_{p-r}\mathbf{S}_{p-r}']'[\mathbf{S}_{p-r}\mathbf{S}_{p-r}']\boldsymbol{\beta} =$$

$$= \boldsymbol{\beta}'\mathbf{S}_{p-r}\mathbf{S}_{p-r}'\mathbf{S}_{p-r}\mathbf{S}_{p-r}'\boldsymbol{\beta}.$$

Însă

$$\mathbf{S}_{p-r}'\mathbf{S}_{p-r} = \mathbf{I}_{p-r}$$

şi deci

$$(Deplasarea)^2 = \boldsymbol{\beta}'\mathbf{S}_{p-r}\mathbf{S}_{p-r}'\boldsymbol{\beta}. \tag{68}$$

Acum

$$\xi_{p-r} = \mathbf{S}_{p-r}'\boldsymbol{\beta}$$

este elementul $(p-r)$ al vectorului de proiecţie al lui $\boldsymbol{\beta}$ pe subspaţiul generat de $\mathbf{S}_{p-r}$. Cu aceste notaţii

$$(Deplasarea)^2 = \xi_{p-r}'\xi_{p-r} = \sum_{j=r+1}^{p} \xi_j^2. \tag{69}$$

Deoarece ξ_j nu depinde de r, avem că $(Deplasarea)^2$ este o funcţie monoton descrescătoare de r. Mai departe,

$(Deplasarea)^2$ are: limita zero pentru $r = p$ şi
valoarea $\boldsymbol{\beta}'\boldsymbol{\beta}$ pentru $r = 0$ (deoarece $\xi_p'\xi_p = \boldsymbol{\beta}'\boldsymbol{\beta}$).

TEOREMA 18 (Marquardt (1970)). *Condiţia suficientă pentru ca eroarea medie pătratică $M(L_1^2)$ să fie mai mică decît covarianţa în* C.M.M.P. *este*

$$\sum_{i=r+1}^{p} \frac{1}{\lambda_j} > \frac{1}{\sigma^2}(\boldsymbol{\beta}'\boldsymbol{\beta}). \tag{70}$$

Demonstraţie. Folosind (66) şi (69), $M(L_1^2)$ este dată de

$$M(L_1^2) = \sigma^2\left(\sum_{j=1}^{r} \frac{1}{\lambda_j}\right) + \sum_{j=r+1}^{p} \xi_j^2$$

în timp ce covarianţa prin C.M.M.P. este

$$\mathscr{C}(\hat{\boldsymbol{\beta}}) = \sigma^2\left(\sum_{j=1}^{p} \frac{1}{\lambda_j}\right).$$

Astfel, condiţia suficientă pentru

$$[M(L_1^2) < \mathscr{C}(\hat{\boldsymbol{\beta}})]$$

este

$$\sigma^2\left(\sum_{j=r+1}^{p}\frac{1}{\lambda_j}\right) > \sum_{j=r+1}^{p}\xi_j^2 \tag{71}$$

sau

$$\sum_{j=r+1}^{p}\left(\frac{\sigma^2}{\lambda_j} - \xi_j^2\right) > 0. \tag{72}$$

Astfel, condiţia suficientă este ca

$$\frac{\sigma^2}{\lambda_j} > \xi_j^2 \qquad \text{pentru toţi } j > r. \tag{73}$$

Deoarece (73) se aplică cu dificultate în practică, o condiţie suficientă, mai puţin puternică, însă mult mai folosită, poate fi obţinută notînd că

$$\sum_{j=1}^{p}\beta_j^2 \geqslant \sum_{j=r+1}^{p}\xi_j^2 \qquad \text{pentru } r \leqslant p. \tag{74}$$

Astfel inegalitatea poate fi scrisă ca în (70).

Marquardt (1970) stabileşte condiţiile în care estimatorul prin inversă generalizată este echivalent cu estimatorul de C.M.M.P.

Observaţii. i) Cu toate că estimatorii ridge şi prin inversă generalizată au multe proprietăţi comune, estimatorul ridge nu este un estimator prin inversă generalizată. De exemplu, inversa ridge nu satisface

$$\mathbf{AA^+A} = \mathbf{A}. \tag{75}$$

Inversa ridge poate fi privită ca o aproximaţie a inversei generalizate.

ii) Soluţia ridge se calculează simplu şi este mai bună în cazul valorilor proprii foarte mici, însă nu zero. Soluţia inversă generalizată este semnificativă în cazul valorilor proprii care sînt precis zero. Această observaţie sugerează folosirea estimatorului combinat

$$\hat{\boldsymbol{\beta}}^{*+} = (\mathbf{A}_r + k\mathbf{I})^{-1}\mathbf{g}, \tag{76}$$

unde r este ales astfel încît să elimine valorile proprii care sînt zero şi k este ales să reducă efectul valorilor proprii apropiate de zero.

§ 3. O clasă de estimatori contractaţi

3.1. Punerea problemei

Reducem modelul (1) la forma canonică
Există o matrice ortogonală $\mathbf{P}$, $n \times n$ şi $\mathbf{Q}$, $p \times p$ astfel încît

$$\mathbf{QX'XQ'} = \Lambda^2 \tag{77}$$

Scanned with OKEN Scanner

şi

$$\mathbf{PXQ'} = \mathbf{D} = \begin{pmatrix} \Lambda \\ \hline 0 \end{pmatrix} \tag{78}$$

unde $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_p\}$.

Scriind

$$\mathbf{Z} = \mathbf{Py}, \quad \boldsymbol{\gamma} = \mathbf{Q}\boldsymbol{\beta}, \quad \mathbf{v} = \mathbf{P}\boldsymbol{\varepsilon},$$

obţinem

$$\mathbf{Z} = \mathbf{D}\boldsymbol{\gamma} + \mathbf{v}. \tag{79}$$

Vom lucra cu presupunerea că $\mathbf{v} \sim N(0, \sigma^2\mathbf{I}_n)$ sau cu alte cuvinte

$$\begin{aligned} &\mathbf{Z}_i \sim N(\lambda_i\gamma_i, \sigma^2), \qquad i = 1, 2, \ldots, p \\ &\mathbf{Z}_i \sim N(0, \sigma^2), \qquad i = p+1, p+2, \ldots, n \end{aligned} \tag{80}$$

Observaţie. Transformările ortogonale păstrează distanţele şi media erorii pătratice a estimatorului $\boldsymbol{\beta}^*$ pentru $\boldsymbol{\beta}$ este egală cu media erorii pătratice a statisticii $\boldsymbol{\gamma}^* = \mathbf{Q}\boldsymbol{\beta}^*$ considerată ca estimator pentru $\boldsymbol{\gamma}$.

Valorile proprii ale matricei $\mathbf{X'X}$ sînt $\{\lambda_1^2, \lambda_2^2, \ldots, \lambda_p^2\}$ şi efectul faptului că cel puţin una dintre λ_i este mică în valoare absolută este *creşterea* valorii absolute a estimaţiei de C.M.M.P.

$$z_i/\lambda_i \text{ pentru } \gamma_i \text{ (Deoarece } M\{(z_i/\lambda_i) - \gamma_i\}^2 = \sigma^2/\lambda_i^2).$$

Pornind de la această observaţie, Goldstein, Smith (1974) examinează posibilitatea *contractării* estimaţiei, considerînd estimatori de forma

$$\gamma_i^* = c_i z_i \text{ cu } |c_i z_i| \leqslant |c_i/\lambda_i|,$$

$$c_i = c(\lambda_i, k), \quad k \geqslant 0.$$

În cazul în care dorim să-l comparăm cu estimaţia de C.M.M.P., luăm $c(\lambda_i, 0) = 1/\lambda_i$, astfel că $k = 0$ corespund la C.M.M.P. Pentru a obţine efectul de *contracţie* vom cere ca $|c(\lambda_i, k)|$ să fie, pentru orice λ_i fixat, o funcţie continuă, monoton descrescătoare de k cînd k creşte de la 0 la ∞. Deasemenea $c(\lambda_i, k)$, pentru toţi k, are acelaşi semn ca λ_i. Aceste condiţii sînt sintetizate în

$$\begin{cases} \text{i)} \ c(\lambda_i, 0) = 1/\lambda_i, \\ \text{ii)} \ c'(\lambda_i, k)/\lambda_i < 0 \quad \text{pentru toţi } k \geqslant 0. \end{cases} \tag{81}$$

LEMA 1. *Pentru orice* $\boldsymbol{\gamma}$, *există* $k > 0$ *astfel* $\gamma_i^* = c(\lambda_i, k)\, z_i$ *are eroarea medie pătratică mai mică decît* $\hat{\gamma}_i = z_i/\lambda_i$ *pentru toţi* $i = 1, 2, \ldots, p$.

Demonstraţie. Punînd $c_i = c(\lambda_i, k)$,

$$M(c_i z_i - \gamma_i)^2 = c_i^2\sigma^2 + \gamma_i^2(c_i\lambda_i - 1)^2$$

şi

$$M\{(z_i/\lambda_i) - \gamma_i\}^2 = \sigma^2/\lambda_i.$$

Urmează că $c_i z_i$ are eroarea medie pătratică mai mică dacă şi numai dacă

$$\gamma_i^2 < \frac{\sigma^2}{\lambda_i^2} \cdot \frac{(1/\lambda_i) + c_i}{(1/\lambda_i) - c_i}, \qquad i = 1, 2, \ldots, p. \tag{82}$$

Proprietăţile presupuse ale lui c_i (date în (81)) asigură existenţa lui k astfel încît (82) este satisfăcută.

TEOREMA 19 (Goldstein, Smith (1974)). *Pentru orice* β, *există* $k > 0$ *astfel încît, pentru* β^* *definit de*

$$\beta^* = \mathbf{Q}'\gamma^*. \quad \gamma_i^* = c(\lambda_i, k)\, z_i, \qquad i = 1, 2, \ldots, p, \tag{83}$$

β_i^* *are eroarea medie pătratică mai mică decît a estimatorului corespunzător* $\hat{\beta}_i$ *de C.M.M.P. pentru toţi* $i = 1, 2, \ldots, p$.

Demonstraţie. Este suficient să arătăm că

$$(\mathrm{d}/\mathrm{d}k)\, M(\beta_i^* - \beta_i)^2|_{k=0} < 0 \text{ pentru toţi } i.$$

Fie q_{ij} elementul (i, j) al lui $\mathbf{Q}$. Avem :

$$M(\beta_i^* - \beta_i)^2 = M\left\{\sum_{j=1}^{p} q_{ji}(\gamma_j^* - \gamma_j)\right\}^2 =$$

$$= \sum_{j=1}^{p} q_{ji}^2 [c^2(\lambda_j, k)\,\sigma^2 + \gamma_j^2 \{c(\lambda_j, k)\lambda_j - 1\}^2] +$$

$$+ \sum_{j=1}^{p} \sum_{h=1}^{p} q_{ji} q_{hi} \gamma_j \gamma_h \{c(\lambda_j, k)\lambda_j - 1\}\, \{c(\lambda_h, k)\, \lambda_h - 1\}.$$

Derivînd în raport cu k, şi evaluînd acest rezultat în $k = 0$, obţinem (din (81))

$$2\sigma^2 \sum_{j=1}^{p} q_{ji}^2 c(\lambda_j, 0)\, c'(\lambda_j, 0) = -2\sigma^2 \sum_{j=1}^{p} q_{ji}^2 \left| \frac{c'(\lambda_j, 0)}{\lambda_j} \right| < 0.$$

3.2. Estimatori ridge şi generalizările lor

Cea mai simplă formă pentru c satisfăcînd (81) este dată de

$$c(\lambda_i, k) = \lambda_i/(\lambda_i^2 + k). \tag{84}$$

Generalizarea constă în posibilitatea diferitelor alegeri ale lui k pentru fiecare componentă.

TEOREMA 20. (Goldstein, Smith (1974). *Dacă* $c(\lambda_i, k)$ *este dat de* (84), *atunci estimatorul* β^* *definit de* (83) *este estimatorul ridge definit de* (12).

Scanned with OKEN Scanner

Demonstrație. În termeni de parametrii ai formei canonice, avem

$$\gamma_i^* = \lambda_i z_i/(\lambda_i^2 + k), \quad i = 1, 2, \ldots, p.$$

Folosind (77) și (78), obținem forma compactă mai complicată

$$\boldsymbol{\gamma}^* = [\mathbf{Q}(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)\,\mathbf{Q}']^{-1}\mathbf{D}'\mathbf{z},$$

și deci, după unele aranjări

$$\boldsymbol{\beta}^* = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}\,\mathbf{X}'\mathbf{y}.$$

Observăm că

$$\frac{\mathrm{d}}{\mathrm{d}k} M(\beta_i^* - \beta_i)^2|_{k=0} = -2\sigma^2 \sum_j q_{ji}^2/\lambda_j^4 \quad (i = 1, 2, \ldots, p).$$

Acesta este cazul cînd β_i are, în spațiul canonic, componente mari în direcția în care estimația este cel mai puțin precisă și prin urmare avem o măsură naturală a modului în care mărimea lui $\mathbf{X}'\mathbf{X}$ afectează estimația pentru β_i individuali. În particular, avem o explicație a cerinței făcute pentru estimațiile ridge și anume că coeficienții cu semne incorecte în $k = 0$ tind la schimbarea la semn corect. Această discuție asupra ratei de descreștere a erorii medii pătratice în zero sugerează o modificare a abordării ridge, care poate fi corespunzătoare dacă sîntem interesați în anumiți β_i. Bazați pe această analiză, putem alege c astfel încît dacă

$$\begin{cases} \sum_j q_{ji}^2/\lambda_j^2 \geqslant c & \text{alegem estimația ridge} \\ \text{în rest} & \text{alegem estimația de C.M.M.P.} \end{cases}$$

Luînd

$$\gamma_i^* = \lambda_i z_i/(\lambda_i^2 + k), \quad i = 1, 2, \ldots, p$$

sîntem conduși la

DEFINIȚIA 3. (Goldstein, Smith (1974)). Estimatorul Hoerl-Kennard generalizat este

(85) $$\boldsymbol{\beta}^* = (\mathbf{X}'\mathbf{X} + \mathbf{QKQ}')^{-1}\mathbf{X}'\mathbf{y},$$

unde $\mathbf{K} = \operatorname{diag}\{k_1, k_2, \ldots, k_p\}$.

Din cele discutate anterior rezultă că există $\{k_1, k_2, \ldots, k_p\}$ astfel încît (85) are media erorii pătrate mai mică decît $\hat{\boldsymbol{\beta}}$.

Media erorii pătratice a lui γ_i^* este

(86) $$M\left(\frac{\lambda_i z_i}{\lambda_i^2 + k_i} - \gamma_i\right)^2 = \frac{\lambda_i^2\sigma^2 + k_i^2\gamma_i^2}{(\lambda_i^2 + k_i)^2}.$$

Ca funcție de k_i, (86) descrește monoton cînd k_i crește de la 0 la σ^2/γ_i^2, și atunci crește monoton (la γ_i^2) cînd k_i crește de la σ^2/γ_i^2 la ∞.

Scanned with OKEN Scanner

Prin urmare (teoretic) avem o alegere optimă pentru fiecare k_i; fiecare k_i fiind în general diferit de ceilalţi.

Goldstein, Smith (1974) sugerează un procedeu iterativ de alegere a lui k_i optim, pornind cu

$$\hat{k}_i = \hat{\sigma}^2/\hat{\gamma}_i^2 = \lambda^2 \sum_{j=p+1}^{n} z_j^2/(n-p)z_i^2. \tag{87}$$

O altă generalizare se obţine luînd

$$c(\lambda_i, k) = \lambda_i^{2m-1}/(\lambda_i^{2m} + k), \qquad i = 1, 2, \ldots, p) \tag{88}$$

unde m este un întreg. (88) satisface (81) şi pentru $m \geqslant 2$ face o analiză mai sensibilă a variaţiei spectrului de valori proprii ($m = 1$ corespunde estimatorului ridge). În termeni de spaţiu parametric original, estimatorul este de forma:

$$\boldsymbol{\beta}^* = [(\mathbf{X}'\mathbf{X})^m + k\mathbf{I}_p]^{-1}(\mathbf{X}'\mathbf{X})^{m-1}\mathbf{X}'\mathbf{Y} \tag{89}$$

3.3. C.M.M.P. cu restricţii

Vom alege $\mathbf{Q}$ în (77) astfel încît

$$\lambda_1^2 \geqslant \lambda_2^2 \geqslant \ldots \geqslant \lambda_p^2$$

şi că $\lambda_{r+1}^2, \lambda_{r+2}^2, \ldots, \lambda_p^2$ sînt presupuse a fi zero. Observăm că estimatorii de C.M.M.P. z_i/λ_i pentru $\gamma_i (i = r+1, \ldots, p)$ nu au sens. În termeni de estimatori contractaţi

$$\lambda_i z_i/(\lambda_i^2 + k_i) \text{ pentru orice } k_i > 0,$$

presupunerea $\lambda_i = 0$ conduce la estimatorul $\gamma_i^* = 0$. În cazul extrem în care $\lambda_1^2, \lambda_2^2, \ldots, \lambda_r^2$ sînt mari şi celelalte presupuse a fi zero, valorile mici ale lui $k_1, k_2, \ldots, k_r$ pot fi judecate ca fiind nesemnificative în estimarea lui $\gamma_1, \gamma_2, \ldots, \gamma_r$ şi se poate alege estimatorul

$$\begin{cases} \gamma_i^* = z_i/\lambda_i & (i = 1, 2, \ldots, r) \\ \gamma_i^* = 0 & (i = r+1, r+2, \ldots, p) \end{cases} \tag{90}$$

Pentru a vedea ce implică aceasta pentru $\boldsymbol{\beta}^*$, procedăm după cum urmează. Partiţionăm $\mathbf{Q}'$

$$\mathbf{Q}' = (\mathbf{Q}_r' : \mathbf{Q}_{p-r}'),$$

unde $\mathbf{Q}_r'$ este $p \times r$ şi $\mathbf{Q}_{p-r}'$ este $p \times (p-r)$ şi definim

$$\Lambda_r^2 = \text{diag}\{\lambda_1^2, \lambda_2^2, \ldots, \lambda_r^2\}.$$

Scanned with OKEN Scanner

Atunci dacă γ^* este definit de (90), β^* definit de $\mathbf{Q}'\gamma^*$ este dat de

$$\beta^* = (\mathbf{X}'\mathbf{X})_r^+ \mathbf{X}'\mathbf{y}, \tag{91}$$

unde $(\mathbf{X}'\mathbf{X})_r^+ = \mathbf{Q}_r' \Lambda_r^{-2} \mathbf{Q}_r$. Ultima este inversa generalizată a lui $\mathbf{X}'\mathbf{X}$ de rang r și (91) este de aceeași formă ca (59).

Dezvoltarea anterioară arată că în abordarea contractată, aceasta corespunde la o împărțire mai crudă în valori proprii „zero" și „mari", implicînd alegerea lui $k = 0$, respectiv $k = \infty$.

La estimatorul (90) se poate ajunge de la estimatorul de C.M.M.P. supus la constrîngerea

$$(\mathbf{0} : \mathbf{I}_{p-r})\,\gamma = \mathbf{0},$$

în presupunerea că $\lambda_i^2 = 0 (i = r+1, r+2, \ldots, p)$. Notînd că constrîngerea nu este alterată dacă înmulțim prin orice constantă arbitrară k nenulă, putem scrie ecuația satisfăcută de γ^* în forma

$$\begin{pmatrix} \Lambda_r^2 & \mathbf{0} \\ \mathbf{0} & k\mathbf{I}_{p-r} \end{pmatrix} \gamma^* = \begin{pmatrix} z_1\lambda_1 \\ \vdots \\ z_r\lambda_r \\ \cdots \\ \mathbf{0} \end{pmatrix}$$

Dacă înlăturăm presupunerea că $\lambda_{r+1} = \lambda_{r+2} = \ldots = \lambda_p = 0$ și le reatribuim valorile lor foarte mici, notăm că dacă alegem $k > 0$ astfel încît

$$k \ll \lambda_i^2 (i = 1, 2, \ldots, r),\ k \gg \lambda_i^2 \quad (i = r+1, r+2, \ldots, p),$$

atunci are loc următoarea aproximare

$$\begin{pmatrix} \Lambda_r^2 + k\mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \Lambda_{p-r}^2 + k\mathbf{I}_{p-r} \end{pmatrix} \gamma^* \simeq \begin{pmatrix} z_1\lambda_1 \\ z_2\lambda_2 \\ \vdots \\ z_p\lambda_p \end{pmatrix}$$

sau

$$(\Lambda^2 + k\mathbf{I}_p)\gamma^* = \mathbf{D}'\mathbf{Z},$$

care în modelul (1) dă estimatorul ridge (12).

Estimatorul ridge poate fi o bună aproximație a estimatorului Marquardt într-un interval rezonabil de alegeri de rang r. Media erorii pătratice a lui (90) este dată de

$$\sum_{i=1}^{r} \sigma^2/\lambda_i^2 + \sum_{i=r+1}^{p} \gamma_i^2,$$

iar a estimatorului ridge (12) (scris în termeni de forma canonică a parametrilor) de :

$$\sum_{i=1}^{p} \left[(\sigma^2/\lambda_i^2) \left\{ \frac{1}{1 + (k/\lambda_i^2)} \right\}^2 + \gamma_i^2 \left\{ \frac{k/\lambda_i^2}{1 + (k/\lambda_i^2)} \right\}^2 \right].$$

Scanned with OKEN Scanner

Dacă $k \ll \lambda_i^2$ $(i = 1, 2, \ldots, r)$ și $k \gg \lambda_i^2$ $(i = r+1, r+2, \ldots, p)$, mediile erorilor pătratice vor fi foarte apropiate. În orice caz, vedem că pentru fiecare componentă există o pondere între σ^2/λ_i^2 și γ_i^2; natura ponderii depinde de raportul lui k la valoarea proprie λ_i^2.

§ 4. Estimatori ridge buni și optimi

4.1. Introducere

Urmînd Obenchain ((1975), (1978)) folosim următoarele notații și convenții:

— $\mathbf{X}$ matricea regresorilor de ordin $n \times p$ și
— $\mathbf{y}$ vectorul corespunzător al răspunsurilor de ordin $n \times 1$
— $\mathbf{1}'\mathbf{X} = \mathbf{0}'$ și $\mathbf{1}'\mathbf{y} = 0$
— modelul standard al regresiei liniare multiple este

$$M(\mathbf{y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \quad \text{și} \quad \mathscr{C}(\mathbf{y}|\mathbf{X}) = \sigma^2(\mathbf{I} - \mathbf{1}\mathbf{1}'/n),$$

unde $\boldsymbol{\beta}$ este vectorul coeficienților regresiei necunoscut de ordin $p \times 1$ $(\boldsymbol{\beta}'\boldsymbol{\beta} < \infty)$ și $\sigma^2 > 0$ este dispersia erorii, necunoscută.
Pentru $\mathbf{X}$ se consideră descompunerea (Rao (1973), p. 42)

$$\mathbf{X} = \mathbf{H}\boldsymbol{\Lambda}^{1/2}\mathbf{G}' \tag{92}$$

unde $\boldsymbol{\Lambda}^{1/2}$ este matricea diagonală a valorilor singulare ordonate ale matricei $\mathbf{X}$, $\lambda_1^{1/2} \geqslant \ldots \geqslant \lambda_p^{1/2}$.
Presupunem că $\lambda_p > 0$ astfel încît $\boldsymbol{\beta}$ este estimabil și fie $\mathbf{b}^0$ estimatorul de C.M.M.P. pentru $\boldsymbol{\beta}$; $\mathbf{b}^0 = \mathbf{X}^+\mathbf{y}$, unde $\mathbf{X}^+$ este inversa Moore-Penrose. Atunci

$$\mathbf{b}^0 = \mathbf{G}\mathbf{c},$$

unde vectorul

$$\mathbf{c} = \boldsymbol{\Lambda}^{-1/2}\mathbf{H}'\mathbf{y}$$

conține componentele necorelate ale lui $\mathbf{b}^0$ și

$$M(\mathbf{c}|\mathbf{X}) = \mathbf{G}'\boldsymbol{\beta} \equiv \boldsymbol{\gamma},$$

componentele adevărate necunoscute ale lui $\boldsymbol{\beta}$.
Elementul i al lui $\mathbf{c}$ are forma:

$$c_i = (\mathbf{y}'\mathbf{y}/\lambda_i)^{-1/2} r_{yi}, \tag{93}$$

unde r_{yi} este coeficientul de corelație dintre y și coordonatele regresorilor în lungul axei principale de ordin i; statistica F în teoria normală pentru $\gamma_i = 0$ este

$$F_i = r_{yi}^2(n - p - 1)/(1 - R^2). \tag{94}$$

unde

$$R^2 = r_{y1}^2 + r_{y2}^2 + \ldots + r_{yp}^2$$

În teoria repartiției condiționate $\mathbf{X}$ (și astfel $\mathbf{H}$) este dată și atunci corelațiile principale $r_{y1}, \ldots, r_{yp}$ nu sînt repartizate ca coeficienții de corelație. Necentricitatea lui F este

$$\phi_i = \gamma_i^2 \lambda_i / \sigma^2,$$

însă din (94) urmează că F_i nu depinde de λ_i.

Folosind $\delta_i c_i$ ca estimator pentru γ_i pentru $i = 1, \ldots, p$, unde δ_i este un factor ridge realizabil (fix sau stocastic). Estimatorul lui $\boldsymbol{\beta}$ este

$$\mathbf{b}^* = \mathbf{G}\boldsymbol{\Delta}\mathbf{c}, \tag{95}$$

unde $\boldsymbol{\Delta}$ este matricea diagonală cu elementele $\delta_1, \ldots, \delta_p$. Dacă dat $\mathbf{X}$, $\delta_1, \ldots, \delta_p$ nu sînt stocastici, atunci matricea mediei erorii pătratice (MMEP) a lui $\mathbf{b}^*$ ca estimatorul a lui $\boldsymbol{\beta}$ este

$$\mathbf{G}[\mathrm{MMEP}(\boldsymbol{\Delta}\mathbf{c})]\mathbf{G}',$$

unde

$$\mathrm{MMEP}(\boldsymbol{\Delta}\mathbf{c}) = \sigma^2\boldsymbol{\Delta}^2\boldsymbol{\Lambda}^{-1} + (\mathbf{I} - \boldsymbol{\Delta})'\boldsymbol{\gamma}\boldsymbol{\gamma}'(\mathbf{I} - \boldsymbol{\Delta}) \tag{96}$$

este matricea mediei erorii pătratice a lui $\boldsymbol{\Delta}\mathbf{c}$ ca estimator a lui $\boldsymbol{\gamma}$.

O măsură ponderată a mediei erorii pătratice (pmep) a lui $\mathbf{b}^*$ ca un estimator a lui $\boldsymbol{\beta}$ este dat de

$$\mathrm{pmep}(\mathbf{b}^*, \mathbf{W}) = M[(\mathbf{b}^* - \boldsymbol{\beta})'\mathbf{W}(\mathbf{b}^* - \boldsymbol{\beta})], \tag{97}$$

unde $\mathbf{W}$ este orice matrice de ponderă nestochastică, nenegativ definită.

Exemplul 1. Pentru $\mathbf{W} = \mathbf{I}$, din (97), obținem cazul considerat de Hoerl, Kennard (1970) și

$$\mathrm{pmep}(\mathbf{b}^*, \mathbf{I}) = \mathrm{Ur}(\mathrm{MMEP}(\mathbf{b}^*))$$

și este numită media erorii pătratice sumată a lui $\mathbf{b}^*$.

Exemplul 2. Dacă $\boldsymbol{\alpha}$ este vectorul unitate ($\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$) atunci

$$\mathrm{pmep}(\mathbf{b}^*, \boldsymbol{\alpha}\boldsymbol{\alpha}') = \boldsymbol{\alpha}'(\mathrm{MMEP}\ (\mathbf{b}^*))\boldsymbol{\alpha} =$$

$$= \mathrm{MMEP}(\boldsymbol{\alpha}'\mathbf{b}^*) = \mathrm{MMEP}(-\boldsymbol{\alpha}'\mathbf{b}^*)$$

este media erorii pătratice a lui $\mathbf{b}^*$ paralelă la $\pm\boldsymbol{\alpha}$ în spațiul euclidian p-dimensional, adică este media erorii pătratice a lui $\boldsymbol{\alpha}'\mathbf{b}^*$ ca un estimator pentru $\boldsymbol{\alpha}'\boldsymbol{\beta}$.

O măsură matricială a mediei erorii pătratice a lui $\mathbf{b}^*$ relativă la $\mathbf{b}^0$ este dată de

$$\mathrm{EMMEP}(\mathbf{b}^*) = \mathrm{MMEP}(\mathbf{b}^0) - \mathrm{MMEP}(\mathbf{b}^*),$$

care va fi numită excesul matricii mediei erorii pătratice (EMMEP).

Observație. $\mathrm{EMMEP}(\mathbf{b}^*)$ nu este $\mathrm{MMEP}(\mathbf{b}^0 - \mathbf{b}^*)$.

Scanned with OKEN Scanner

DEFINIȚIA 4 (Obenchain (1978)). $\mathbf{b}^*$ din (95) poartă numele de estimator ridge generalizat *bun* pentru $\boldsymbol{\beta}$ dacă următoarele trei proprietăți echivalente au loc:

(98 i) $$\text{EMMEP}(\mathbf{b}^*) = \text{MMEP}(\mathbf{b}^0) - \text{MMEP}(\mathbf{b}^*)$$

este o matrice pozitiv definită, sau

(98 ii) $$\text{MMEP}(\boldsymbol{\alpha}'\mathbf{b}^0) > \text{MMEP}(\boldsymbol{\alpha}'\mathbf{b}^*)$$

pentru fiecare vector unitate $\boldsymbol{\alpha}$, sau

(98 iii) $$\text{EMMEP}(\mathbf{b}^0, \mathbf{W}) > \text{EMMEP}(\mathbf{b}^*, \mathbf{W})$$

pentru fiecare matrice de pondere $\mathbf{W}$ pozitiv definită.

Observație. Theobald (1974) în teorema 1 demonstrează echivalența proprietăților (98 i) și (98 iii).

4.2. Factori ridge buni

Fie $\zeta_1 \geqslant \ldots \geqslant \zeta_p$ valorile proprii ordonate ale EMMEP $(\mathbf{b}^*)$ (98 i). Atunci $\mathbf{b}^*$ este bun dacă și numai dacă $\zeta_p > 0$. Definim o funcție cu valori scalari $RF(\Delta)$ de $\delta_1, \ldots, \delta_p$ prin

$$RF(\Delta) = \sum_{i=1}^{p} [\delta_i^{\text{MMEP}}(1 - \delta_i)]/[(1 - \delta_i^{\text{MMEP}})(1 + \delta_i)] =$$

(99)

$$= \sum_{i=1}^{p} \phi_i^2(1 - \delta_i)/(1 + \delta_i),$$

unde

$$\delta_i^{\text{MMEP}} = \phi_i^2/(1 + \phi_i^2),$$

și $\phi_i^2 = \gamma_i^2\lambda_i/\sigma^2$ este din nou necentricitatea necunoscută a lui F_i (94). Următoarea teoremă dă condiții necesare și suficiente pentru ca un oarecare $\mathbf{b}^*$ să fie bun și reliefează vectorul propriu $\boldsymbol{\alpha}^*$ al lui EMMEP $(\mathbf{b}^*)$ ce corespunde la ζ_p cînd $\zeta_p < 0$.

TEOREMA 21. (*Teorema funcției ridge*) (Obenchain (1978)). *Dacă* $(\boldsymbol{\beta}, \sigma^2)$ *sînt parametrii fixați astfel încît* $\sigma^2 > 0$ *și* $\boldsymbol{\beta}'\boldsymbol{\beta} < \infty$, $\mathbf{X}$ *dată este de rang* $p(\lambda_p > 0)$ *și* δ_i *nu sînt stocastici,* $0 \leqslant \delta_i \leqslant 1$ *pentru* $i = 1, \ldots, p$, *atunci*

i) $\zeta_{p-1} > 0$ *dacă* $p > 1$

ii) $\zeta_p > 0$ *dacă și numai dacă toți* $\delta_1, \ldots, \delta_p$ *sînt suficient de apropiate de unu astfel încît* $RF(\Delta) < 1$,

iii) $\zeta_p = 0$ *dacă și numai dacă* $RF(\Delta) = 1$,

iv) $\zeta_p < 0$ *dacă și numai dacă* $RF(\Delta) > 1$ și

Scanned with OKEN Scanner

v) *vectorul propriu al* EMMEP ($\mathbf{b}^*$) *corespunzător la* $\zeta_p < 0$ *are elementul* j

$$\alpha_j^* \simeq \sum_{i=1}^{p} \frac{g_{ji}(1-\delta_i)\gamma_i}{[\sigma^2\lambda_i^{-1}(1-\delta_i^2)+|\zeta_p|]} \tag{100}$$

pentru $\mathbf{G} = ((g_{ji}))$ *şi* λ_i *din* (92) şi $\boldsymbol{\gamma} = \mathbf{G}'\boldsymbol{\beta}$.

Observaţii. i) În cazul în care

$$\delta_i = \lambda_i/(\lambda_i + k),$$

funcţia ridge este

$$RF(\Delta) = \sum \phi_i^2/(1 + 2\lambda_i k^{-1}).$$

Fiecare $k > 0$ dă un estimator bun dacă $\sum \phi_i^2 < 1$; în rest $0 < k < 2/|\eta_p|$ unde η_p este valoarea proprie negativă a lui $(\mathbf{X}'\mathbf{X})^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}'/\sigma^2$ este un interval bun. Astfel condiţia suficientă a lui Theobald (1974) din Teorema 2, ca $0 < k < 2\sigma^2/\boldsymbol{\beta}'\boldsymbol{\beta}$ este mai stringentă decît este necesar;

ii) Cînd $p = 1$, $\delta_1 c_1$ este bun pentru $\delta_1 > \delta_1^{\text{MIN}} = \max(0; 2\delta_1^{\text{MMEP}} - 1)$ şi $RF(\Delta^{\text{MMEP}}) < 1$ cînd $p = 2$. $RF(\Delta^{\text{MMEP}})$ poate depăşi 1 cînd $p > 2$. $\delta_i \geqslant 1 - 2(1 - \delta_i^{\text{MMEP}})/p$ pentru $i = 1, \ldots, p$ este suficientă pentru ca $\mathbf{b}^*$ să fie bun deoarece fiecare termen din (99) va fi atunci mai mic decît $1/p$.

iii) Deoarece $f(\delta) = (1 - \delta)/(1 + \delta)$ este convexă pe $0 \leqslant \delta \leqslant 1$, factorii ridge buni formează o mulţime convexă, deschisă, nevidă în interiorul lui $0 \leqslant \delta_i < 1$ pentru $i = 1, \ldots, p$ pentru fiecare $(\boldsymbol{\beta}, \sigma^2)$ fixat cu $\boldsymbol{\beta}'\boldsymbol{\beta} < \infty$ şi $\sigma^2 > 0$.

iv) Dacă $RF_1(\Delta)$ este valoarea funcţiei ridge în $(\boldsymbol{\beta}_1, \sigma_1^2)$ şi f este orice factor constant nenul, atunci $RF_2(\Delta) = f^2 RF_1(\Delta)$ este valoarea sa în $(\boldsymbol{\beta}_2, \sigma_2^2)$ de forma $(f\boldsymbol{\beta}_1, \sigma_1^2)$ sau $(\boldsymbol{\beta}_1, \sigma_1^2/f^2)$. Deoarece $RF_1(\Delta)$ trebuie să fie > 0 pentru unii $(\boldsymbol{\beta}_1, \sigma_1^2)$ cînd δ_i nu este stocastic şi $0 \leqslant \delta_i < 1$ pentru $i = 1, \ldots, p$, nici un astfel de Δ nu poate fi bun pentru toţi $(\boldsymbol{\beta}, \sigma^2)$ cu $\boldsymbol{\beta}'\boldsymbol{\beta} < \infty$ şi $\sigma^2 > 0$.

Estimaţie în teoria normală. Estimatorul de verosimilitate maximă pentru $\phi_i^2 = \gamma_i^2\lambda_i/\sigma^2$ este $nF_i/(n - p - 1)$ pentru F_i din (94). Estimatorii de verosimilitate maximă pentru $\zeta_1, \ldots, \zeta_p$ sînt cele p soluţii ale

$$\sum_{i=1}^{p} c_i^2(1 - \delta_i)^2/[s^2\lambda_i^{-1}(1 - \delta_i^2) - \zeta] = 1 \tag{101}$$

dacă $c_i = (\mathbf{y}'\mathbf{y}/\lambda_i)^{1/2} r_{yi}$ nu este zero pentru $i = 1, \ldots, p$ şi $s^2 = (\mathbf{y}'\mathbf{y})(1 - R^2)/n$. Dacă estimaţia lui ζ_p este negativă, înlocuirea în (94) dă estimatorul de verosimilitate maximă ce corespunde la $\boldsymbol{\alpha}^*$.

Estimatorul lui ϕ_i^2 care este ndeeplasat în teoria normală este

$$\hat{\phi}_i^2 = -1 + (n - p - 3)F_i/(n - p - 1)$$

şi estimatorul *interval corect* corespunzător este

$$\hat{\phi}_i^{*2} = \max(0, \hat{\phi}_i^2).$$

Scanned with OKEN Scanner

Toţi estimatorii anteriori pot fi folosiţi pentru a estima

$$RF(\Delta) = \sum \phi_i^2(1 - \delta_i)/(1 + \delta_i).$$

Matricea mediei erorii pătratice reduse, $\mathrm{MMEP}(\Delta\mathbf{c})/\sigma^2$ este estimată nedeplasat pentru Δ nealeator de

$$\hat{\mathbf{T}} = \frac{(n-p-3)}{(n-p-1)}(\mathbf{I}-\Delta)\mathbf{A}^{-1/2}\mathbf{t}\mathbf{t}'\mathbf{A}^{-1/2}(\mathbf{I}-\Delta) + \tag{102}$$

$$+\mathbf{A}^{-1}(2\Delta - \mathbf{I}) = ((\hat{\tau}_{ij})),$$

unde $\mathbf{t}$ este vectorul coloană al statisticilor $t_i = \triangleq F_i^{1/2}$ din (94). Elementul diagonală de ordin i al $\mathrm{MMEP}(\Delta\mathbf{c})/\sigma^2$ este

$$\tau_{ii} = \mathrm{MMEP}(\delta_i c_i)/\sigma^2 = [(1-\delta_i)^2\phi_i^2 + \delta_i^2]/\lambda_i \geqslant \delta_i^2/\lambda_i.$$

Estimatorii lui $\mathrm{MMEP}(\alpha'\ \mathbf{b}^*)/\sigma^2$ sînt $\alpha'\mathbf{G}\hat{\mathbf{T}}\mathbf{G}'\alpha$ pentru $\mathbf{G}$ din (92).

Cazurile particulare de interes apar cînd α este o coloană a matricei identitate sau cînd α indică un contrast între perechi de coeficienţi despre care se ştie că sînt puternic negativ corelaţi în estimatorul de C.M.M.P.

În restricţia $\Delta = \delta\mathbf{I}$, $\mathbf{b}^* = \delta\mathbf{b}^0$ este bun pentru $\delta > 1 - 2/(1 + \sum\phi_i^2)$ unde $\sum\phi_i^2$ este necentricitatea raportului F,

$$F_0 = (F_1 + \ldots + F_p)/p = R^2(n-p-1)/[p(1-R^2)]$$

pentru verificarea ipotezei $\beta = 0$.

4.3. Factori ridge optimi

Următoarele teoreme caracterizează factorii ridge optimi.

DEFINIŢIA 5. Factorii ridge care ating fie minimum mediei erorii pătratice paralel cu o direcţie arbitrară α în spaţiul coeficienţilor sau minimum $\mathrm{pmep}(\mathbf{b}^*, \mathbf{W})$, pentru o matrice de pondere $\mathbf{W}$ arbitrară, pozitiv definită poartă numele de factori ridge optimi.

Vom arăta că direcţiile paralele şi ortogonale cu β necunoscut joacă un rol pivotal în această teorie.

TEOREMA 22 (Obenchain (1978)). *Dat vectorul unitate* α, $\sigma^2 > 0$ *şi* $\beta'\beta < \infty$,

i) $\mathrm{MMEP}(\alpha'\mathbf{b}^*)$ *nu depinde de* δ_i *cînd* $\alpha'g_i = 0$, şi

ii) $\mathrm{MMEP}(\alpha'\mathbf{b}^*)$ *este minimizată prin alegerea lui* δ_i *nealeator*

$$\delta_i(\alpha) = (\alpha'\beta)\lambda_i\gamma_i/[(\alpha'\mathbf{g}_i)(\sigma^2 + \Sigma^*\gamma_j^2\lambda_i)] \tag{103}$$

cînd $\alpha'\mathbf{g}_i \neq 0$, *unde* Σ^* *reprezintă suma numai după acei* j *astfel încît* $\alpha'\mathbf{g}_j = 0$. *Echivalent*

$$\delta_i\gamma_i = k\phi_i^2/(\alpha'\mathbf{g}_i) \text{ cînd } k = (\alpha'\beta)(1 + \Sigma^*\phi_j^2).$$

Scanned with OKEN Scanner

Demonstrație. Fie $\xi' = \alpha'G$ vectorul unitate ce dă coordonatele lui α în raport cu axele principale ale lui $\mathbf{X}$. Atunci

$$\alpha'\mathbf{b}^* = \xi'\Delta c$$

este componenta lui $\mathbf{b}^*$ în direcția α și aceasta nu depinde de δ_i cînd α este ortogonal la axa principală $i(\xi_i = \alpha'g_i = 0)$.
Faptul că δ_i nu sînt stocastice implică

$$\mathrm{MMEP}(\alpha'\mathbf{b}^*) = \sigma^2\xi'\Delta^2\Lambda^{-1}\xi + (\xi'(\mathbf{I} - \Delta)\gamma)^2 \text{ și}$$

$$\partial\mathrm{MMEP}(\alpha'\,\mathbf{b}^*)/\partial\delta_i = 2\delta_i\xi_i^2\lambda_i^{-1}\sigma^2 - 2(\xi'(\mathbf{I} - \Delta)\,\gamma)\xi_i\gamma_i,$$

care este identic zero cînd $\xi_i = 0$ sau devine zero cînd

$$\delta_i = k^{(\alpha)}\gamma_i\lambda_i/\xi_i \text{ și } k^{(\alpha)} = [\xi'(\mathbf{I} - \Delta)\gamma]/\sigma^2 \text{ pentru } \xi_i = 0.$$

Problema de minimizare este convexă deoarece matricea derivatelor de ordinul doi este negativ definită; astfel (103) rezultă rezolvînd:

$$\sigma^2k^{(\alpha)} = \xi'\gamma - k^{(\alpha)}\Sigma^*\gamma_j^2\lambda_j$$

pentru $k^{(\alpha)}$ și notînd că $\xi'\gamma = \alpha'\beta$.

Observații. i) $\delta_i(\alpha) = \delta_i(-\alpha)$

$$\text{ii) } \delta_i(g_i) = \gamma_i^2\lambda_i/(\sigma^2 + \gamma_i^2\lambda_i) = \delta_i^{\mathrm{MMEP}}$$

și $\delta_j(g_i)$ este nedeterminată pentru $j \neq i$.

iii) Dacă β este zero, fiecare α este ortogonal cu β în sensul că $\alpha'\beta = 0$, și în acest caz $\delta_i(\alpha) = 0$ este o alegere optimă pentru $i = 1, \ldots$ $\ldots, p$ și pentru fiecare α.

Cînd $\beta \neq 0$, există un spațiu $(p-1)$ dimensional al lui α ortogonal lui β și $\delta_i(\alpha) = 0$ cînd $\alpha'g_i \neq 0$ este din nou o alegere optimă pentru acest α ortogonal.

iv) Dacă $\beta \neq 0$, α paralel cu β dă

$$\xi_i = \alpha'g_i = \gamma_i(\gamma'\gamma)^{-1/2}.$$

Astfel $\delta_i(\alpha)$ este determinat pentru acest α cînd $\gamma_i = 0$ și $\delta_i = k^{(-)}\lambda_i$ este în rest optim pentru

$$k^{(-)} = (\gamma'\gamma)/(\sigma^2 + \gamma'\Lambda\gamma) \geq 0. \tag{104}$$

Estimatorul ridge corespunzător este

$$\mathbf{b}^* = k^{(-)}G\Lambda c = k^{(-)}\mathbf{X}'\mathbf{y}$$

unde $\mathbf{X}'\mathbf{y}$ este vectorul a p produse interioare a valorilor regresorilor cu vectorul răspuns.

Scanned with OKEN Scanner

TEOREMA 23 (Obenchain (1978)). *Dată matricea de ponderе* $\mathbf{W}$, $\sigma^2 > 0$, *şi* $\boldsymbol{\beta}'\boldsymbol{\beta} < \infty$, pmep($\mathbf{b}^*$, $\mathbf{W}$) *din* (97) *este minimizată prin alegerea lui* $\delta_1, \ldots, \delta_p$ *nestocastici*

$$\delta_i(\mathbf{W}) = \gamma_i \lambda_i \eta_i / (\sigma^2 m_{ii}), \tag{105}$$

unde $\boldsymbol{\eta} = (\mathbf{D} + \mathbf{M}^{-1})^{-1}\boldsymbol{\gamma}$, $\mathbf{M} = ((m_{ij})) = \mathbf{G}'\mathbf{W}\mathbf{G}$ şi $\mathbf{D}$ *este matricea diagonală cu elementul* i, ϕ_i^2/m_{ii}. *Echivalent*

$$\boldsymbol{\Delta\gamma} = \mathbf{D}\boldsymbol{\eta} = \mathbf{D}(\mathbf{D} + \mathbf{M}^{-1})^{-1}\boldsymbol{\gamma}.$$

Demonstraţie

$$\text{pmep}(\mathbf{b}^*, \mathbf{W}) = \sigma^2 \text{Ur}(\mathbf{M}\boldsymbol{\Delta}^2\boldsymbol{\Lambda}^{-1}) + \boldsymbol{\gamma}'(\mathbf{I} - \boldsymbol{\Delta})\mathbf{M}(\mathbf{I} - \boldsymbol{\Delta})\boldsymbol{\gamma}$$

este o funcţie convexă de $\delta_1, \ldots, \delta_p$ astfel că minimul se atinge în $\partial\text{pmep}/\partial\boldsymbol{\delta} = 0$, care sînt punctele fixe ale ecuaţiilor de forma (105) cu $\boldsymbol{\eta} = \mathbf{M}(\mathbf{I} - \boldsymbol{\Delta})\boldsymbol{\gamma}$. Astfel

$$\boldsymbol{\eta} = \mathbf{M}\boldsymbol{\gamma} - \mathbf{M}\mathbf{D}\boldsymbol{\eta}$$

dă rezultatul dorit cînd o rezolvăm pentru $\boldsymbol{\eta}$ deoarece

$$(\mathbf{M}\mathbf{D} + \mathbf{I})^{-1} = (\mathbf{D} + \mathbf{M}^{-1})^{-1}\mathbf{M}^{-1}$$

unde $\mathbf{M}$ este pozitiv definită.

Estimaţia în teoria normală. Estimatorul de verosimilitate maximă a lui $k^{(-)}$ dat de (104) este

$$\hat{k}^{(-)} = \sum nr_{yi}^2 \lambda_i^{-1} / [1 + (n-1)R^2].$$

Înlocuind în această formulă n prin $(n - p - 1)$ vom înlătura deplasarea în estimaţia de verosimilitate maximă pentru σ^2, însă nu vom produce un estimator nedeplasat pentru $k^{(-)}$. $\hat{k}^{(-)}$ este sensibil la valorile proprii mici $\lambda_p, \lambda_{p-1}, \ldots$. Estimaţiile (103) şi (105) pot fi construite folosind c_i ca estimaţii de verosimilitate maximă, nedeplasate pentru γ_i şi estimaţiile lui ϕ_i^2 descrise anterior.

§ 5. Validarea mixtă pentru alegerea unui parametru ridge bun

Fie familia de estimatori ridge $\hat{\boldsymbol{\beta}}(\lambda)$ daţi de

$$\hat{\boldsymbol{\beta}}(\lambda) = (\mathbf{X}'\mathbf{X} + n\lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}. \tag{106}$$

Estimaţia $\hat{\boldsymbol{\beta}}(\lambda)$ este media posteriori a lui $\boldsymbol{\beta}$ dacă $\boldsymbol{\beta}$ are repartiţia priorică $N(0, a\mathbf{I})$ şi $\lambda = \sigma^2/na$. $\hat{\boldsymbol{\beta}}(\lambda)$ este deasemeea soluţia problemei:

Să se găsească $\boldsymbol{\beta}$ care satisface constrîngerea

$$\|\boldsymbol{\beta}\| = \gamma$$

Scanned with OKEN Scanner

şi pentru care

$$\frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\| = \min.$$

Aici $\|\ \|$ indică norma euclidiană. Introducînd lagrangianul găsim că problema anterioară este echivalentă cu a găsi minimul după $\boldsymbol{\beta}$ al expresiei :

$$\frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \tag{107}$$

unde λ este multiplicatorul lui Lagrange.

DEFINIŢIA 6. Metoda de minimizare a ecuaţiei (107), în teoria aproximării, poartă numele de metoda regularizării (Hilgers (1976), Wahba (1977)).

Fie $\beta^{(k)}(\lambda)$ estimatorul ridge (106) cu y_k omis. În acest mod dacă λ este o bună alegere, atunci componenta k : $[\mathbf{X}\boldsymbol{\beta}^{(k)}(\lambda)]_k$ a lui $\mathbf{X}\boldsymbol{\beta}^{(k)}(\lambda)$ va fi un predictor bun pentru y_k.

Are loc

DEFINIŢIA 7 (Allen 1971)). Statistica care minimizează

$$P(\lambda) = \frac{1}{n}\sum_{k=1}^{n}([\mathbf{X}\boldsymbol{\beta}^{(k)}(\lambda)]_k - y_k)^2 \tag{108}$$

poartă numele de estimatorul validare mixtă pentru λ.

După formula Sherman-Morrison—Woodbury (a se vedea Housenholder (1964)),

$$P(\lambda) = \frac{1}{2}\|\mathbf{B}(\lambda)(\mathbf{I} - \mathbf{A})\mathbf{y}\|^2 \tag{109}$$

unde $\mathbf{B}(\lambda)$ este matricea diagonală cu elementul (j, j)

$$1/(1 - a_{jj}(\lambda)),$$

$a_{jj}(\lambda)$ fiind elementul (j, j) al matricii

$$\mathbf{A}(\lambda) = \mathbf{X}(\mathbf{X}'\mathbf{X} + n\lambda\mathbf{I})^{-1}\mathbf{X}'.$$

Observaţie. Cînd intrările lui $\mathbf{X}$ sînt 0 cu excepţia intrărilor x_{ii}, $i = 1, 2, \ldots, p$, atunci $[\mathbf{X}\boldsymbol{\beta}^{(k)}(\lambda)]_k$ nu poate fi un bun predictor pentru y_k. În fapt, în acest caz $\mathbf{A}(\lambda)$ este diagonală

$$P(\lambda) = \frac{1}{n}\sum_{k=1}^{n} y_k^2,$$

şi astfel $P(\lambda)$ nu are un minim unic. Urmează că estimatorul validare mixtă nu va fi foarte bun în cazul diagonal.

Scanned with OKEN Scanner

DEFINIȚIA 8 (Golub, Heath, Wahba (1979)). Statistica care minimizează $V(\lambda)$ dat de

$$V(\lambda) = \frac{1}{n} \| (\mathbf{I} - \mathbf{A}(\lambda))\mathbf{y} \|^2 / \left[\frac{1}{2} \mathrm{Ur}(\mathbf{I} - \mathbf{A}(\lambda)) \right]^2 \tag{110}$$

unde

$$\mathbf{A}(\lambda) = \mathbf{X}(\mathbf{X}'\mathbf{X} + n\lambda\mathbf{I})^{-1}\mathbf{X}' \tag{111}$$

poartă numele de estimație validare mixtă generalizată.

TEOREMA 24 (Golub, Heath, Wahba (1979)). *Fie*

$$\mu_1 = \frac{1}{n} \mathrm{Ur}\mathbf{A}(\lambda)$$

$$\mu_2 = \frac{1}{n} \mathrm{Ur}\mathbf{A}^2(\lambda), \quad \mathbf{b}^2 = \frac{1}{n} \| (\mathbf{I} - \mathbf{A}(\lambda))\mathbf{g} \|^2, \quad \mathbf{g} = \mathbf{X}\boldsymbol{\beta}.$$

Atunci

$$\frac{MT(\lambda) - MV(\lambda) + \sigma^2}{MT(\lambda)} = \frac{-\mu_1(2 - \mu_1)}{(1 - \mu_1)^2} + \frac{\sigma^2}{b^2 + \sigma^2\mu_2} \cdot \frac{\mu_1^2}{(1 - \mu_1)^2} \tag{112}$$

și astfel

$$\frac{|MT(\lambda) - MV(\lambda) + \sigma^2|}{MT(\lambda)} < \left(2\mu_1 + \frac{\mu_1^2}{\mu_2} \right) \frac{1}{(1 - \mu_1)^2},$$

$$0 < \mu_1 < 1, \quad T(\lambda) = \frac{1}{n} \| \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda) \|^2.$$

Demonstrație. Deoarece

$$MT = b^2 + \sigma^2\mu_2$$

$$MV = [b^2 + \sigma^2(1 - 2\mu_1 + \mu_2)]/(1 - \mu_2)^2,$$

rezultatul urmează din

$$MT - MV = (b^2 + \sigma^2\mu_2) \left(1 - \frac{1}{(1 - \mu_1)^2} \right) - \sigma^2 \frac{1 - 2\mu_1}{(1 - \mu_1)^2}$$

și deci

$$MT + \sigma^2 - MV = MT \left(1 - \frac{1}{(1 - \mu_1)^2} \right) + \sigma^2 \frac{\mu_2}{(1 - \mu_1)^2}.$$

Observație. Această teoremă implică că dacă

$$\frac{1}{n} \mathrm{Ur}\mathbf{A}(\lambda) = \mu_1 \to 0 \quad \text{cînd } n \to \infty$$

Scanned with OKEN Scanner

şi

$$\left(\frac{1}{n}\,\mathrm{UrA}(\lambda)\right)^2 \Big/ \left(\frac{1}{n}\,\mathrm{UrA}^2(\lambda)\right) = \frac{\mu_1^2}{\mu_2} \to 0 \text{ cînd } n \to \infty,$$

atunci diferenţa dintre

$$MT(\lambda) + \sigma^2 \text{ şi } MV(\lambda)$$

este mică comparativ cu $MT(\lambda)$. Acest rezultat şi faptul că în cazul diagonal extrem $P(\lambda)$ nu are un minim unic sugerează că valoarea care realizează minimul lui $V(\lambda)$ este preferabilă valorii care realizează minimul lui $P(\lambda)$ dacă dorim să alegem λ să minimizeze

$$\frac{1}{n}\,\mathrm{M}_{y^*}\,\|y^* - \mathbf{X}\boldsymbol{\beta}(\lambda)\|^2.$$

COROLARUL 6. *Fie*

$$h = \left(2\mu_1 + \frac{\mu_1^2}{\mu_2}\right)\frac{1}{(1-\mu_1)^2}.$$

Fie λ^0 *valoarea care realizează minimul lui* $MT(\lambda)$. *Atunci* $MV(\lambda)$ *are întotdeauna un minim (posibil local)* $\tilde{\lambda}$ *astfel încît, „ineficienţa medie"* I^0 *definită de*

$$I^\circ = \frac{MT(\tilde{\lambda})}{MV(\lambda^0)}$$

satisface

$$I^\circ \leqslant \frac{1+h(\lambda^0)}{1-h(\tilde{\lambda})}.$$

Demonstraţie. Fie

$$\Lambda = \{\lambda,\ 0 \leqslant \lambda \leqslant \infty,\ MV(\lambda) - \sigma^2 \leqslant T(\lambda^0)(1+h(\lambda^\circ))\}.$$

Deoarece

$$MT(\lambda)(1-h(\lambda)) < MV(\lambda) - \sigma^2 < MT(\lambda)\,(1+h(\lambda)),\quad 0 \leqslant \lambda < \infty,$$

şi MT, MV şi h sînt funcţii continui de λ, atunci Λ este o mulţime închisă nevidă. Dacă 0 nu este un punct de frontieră a lui Λ, atunci $\mathrm{MV}(\lambda) - \sigma^2$ are cel puţin un minim ($\tilde{\lambda}$) în interiorul lui Λ (fig. 4). Acum din teorema 24

$$MT(\tilde{\lambda})(1-h(\tilde{\lambda})) < MV(\tilde{\lambda}) - \sigma^2 < MT(\lambda^\circ)(1+h(\lambda^\circ))$$

şi astfel

$$I^0 = \frac{T(\tilde{\lambda})}{T(\lambda^0)} \leqslant \frac{1+h(\lambda^0)}{1-h(\tilde{\lambda})}.$$

Dacă Λ include pe 0, atunci $\tilde{\lambda}$ poate fi pe frontiera lui Λ, adică $\tilde{\lambda} = 0$, însă marginea anterioară pentru I° are loc.

Scanned with OKEN Scanner

Exemplul 3. Avem

$$\mu_1 = \frac{1}{n}\,\mathrm{UrA} = \frac{1}{n}\sum_{\nu=1}^{p}\frac{\lambda_{\nu n}}{\lambda_{\nu n}+n\lambda} \leqslant \frac{p}{n}$$

$$\frac{\mu_1^2}{\mu_2} = \frac{\left(\frac{1}{n}\,\mathrm{UrA}\right)^2}{\frac{1}{n}\,\mathrm{UrA}^2} = \frac{1}{n}\,\frac{\left(\sum_{\nu=1}^{p}\frac{\lambda_{\nu n}}{\lambda_{\nu n}+\lambda}\right)^2}{\sum_{\nu=1}^{p}\left(\frac{\lambda_{\nu n}}{\lambda_{\nu n}+\lambda}\right)^2} \leqslant \frac{p}{n}$$

unde $\lambda_{\nu n}$, $\nu = 1, 2, \ldots, n$ sînt valorile proprii ale lui $\mathbf{XX}'$, $\lambda_{pn} = 0$, $\nu > p$.

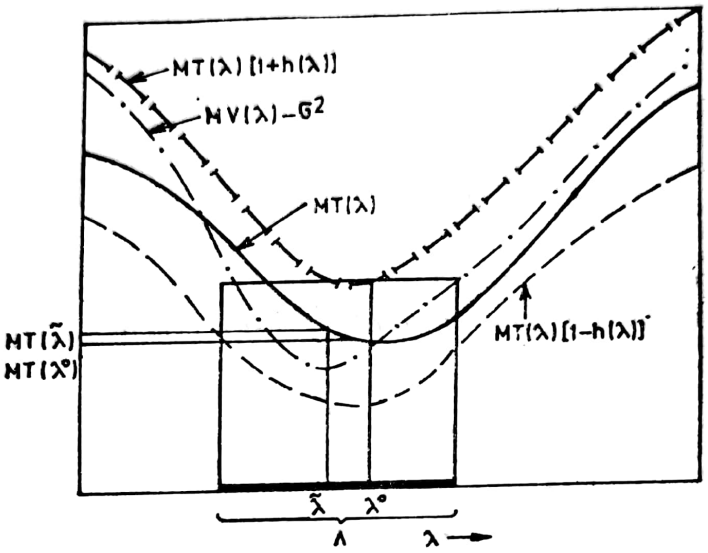


Fig. 4

Atunci

$$h \leqslant 3\,\frac{p}{n}\cdot\frac{1}{\left(1-\frac{p}{n}\right)^2}.$$

Deci pentru p fixat şi $n \to \infty$, urmează că

$$I^0 \leqslant 1 + 6\,\frac{p}{n} + 0\left(\frac{p}{n}\right).$$

Exemplul 4. $p > n$.

Nu este necesar ca $p \ll n$ pentru ca I^0 să tindă la 1, cum sugerează exemplul 1. Este cerut ca $\mathbf{XX}'$ să aibă valori proprii mici pentru n mare.

Fie

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i, \quad i = 1, 2, \ldots, p > n \tag{113}$$

cu

$$\sum_{j=1}^{\infty} x_{ij}^2 \leqslant k_1 < \infty, \text{ pentru toţi } i, \quad \sum_{j=1}^{\infty}\beta_j^2 \leqslant k_2 < \infty.$$

Scanned with OKEN Scanner

Presupunem că

$$\lim_{n\to\infty}\frac{1}{n}\operatorname{Ur}\mathbf{XX}'=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{\infty}x_{ij}^2=k_3<\infty$$

şi presupunem că valorile proprii $\{\lambda_{\nu n},\ \nu=1,2,\ldots,n\}$ ale lui $\mathbf{XX}'$ satisfac

$$\lambda_{\nu n}\simeq n\nu^{-m},\ \text{pentru}\ m>1\left(k_3=\sum_{\nu=1}^{\infty}\nu^{-m}\right).$$

Atunci

$$\mu_1=\frac{1}{n}\sum_{\nu=1}^{n}\frac{\lambda_{\nu n}}{\lambda_{\nu n}+n\lambda}\simeq\frac{1}{n}\sum_{\nu=1}^{n}\frac{1}{1+\lambda\nu^m}\simeq$$

$$\simeq\frac{1}{n}\int_0^{\infty}\frac{\mathrm{d}x}{1+\lambda x^m}=\frac{1}{n\lambda^{1/m}}\int_0^{\infty}\frac{\mathrm{d}x}{1+x^m},$$

$$\mu_2=\frac{1}{n}\sum_{\nu=1}^{n}\left(\frac{\lambda_{\nu n}}{\lambda_{\nu n}+n\lambda}\right)^2\simeq\frac{1}{n}\sum_{\nu=1}^{n}\frac{1}{(1+\lambda\nu^m)^2}\simeq$$

$$\simeq\frac{1}{n}\int_0^{\infty}\frac{\mathrm{d}x}{(1+\lambda x)}=\frac{1}{n\lambda^{1/m}}\int_0^{\infty}\frac{\mathrm{d}x}{(1+x^m)^2}$$

şi dacă $n\lambda^{1/m}\to\infty$, $\mu_1\to 0$, $\mu_1^2/\mu_2\to 0$.

Deoarece cea mai mare valoare proprie a

$$(\mathbf{X}'\mathbf{X}+n\lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X}+n\lambda\mathbf{I})^{-1}=\max_{\nu}\frac{(\lambda_{\nu n})(n\lambda)}{(\lambda_{\nu n})^2+(n\lambda)^2}\leqslant 1$$

urmează că

$$b^2(\lambda)=\lambda\boldsymbol{\beta}'(\mathbf{X}'\mathbf{X}+n\lambda\mathbf{I})^{-1}(n\lambda)\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X}+n\lambda\mathbf{I})^{-1}\boldsymbol{\beta}\leqslant\frac{\lambda}{2}\|\boldsymbol{\beta}\|^2\leqslant\frac{\lambda}{2}k_2.$$

Cînd $n\to\infty$, şirul $\lambda^0=\lambda^0(n)$ ce minimizează $MT(\lambda)=b^2(\lambda)+\sigma^2\mu_2(\lambda)$ trebuie să satisfacă $\lambda^0\to 0$, $n(\lambda^0)^{1/m}\to\infty$, astfel încît poate fi aplicată teorema 24.

În loc să privim $\boldsymbol{\beta}$ fix însă necunoscut, presupunem că $\boldsymbol{\beta}$ ar repartiţia aprioric㠄 $\boldsymbol{\beta}\sim N(\mathbf{0},a\mathbf{I})$. Fie $M_{\boldsymbol{\beta}}$ media în raport cu repartiţia aprioric㠄 şi M media în raport cu ε. Atunci are loc

Teorema 25 (Golub, Heath, Wahba (1979)). *Punctul minim al funcţiei* $M_{\boldsymbol{\beta}}MV(\lambda)$ *este acelaşi ca punctul de minim al funcţiei* $M_{\boldsymbol{\beta}}MT(\lambda)$ *şi este* $\lambda=\sigma^2/na$.

Demonstraţie. Deoarece $M\mathbf{gg}'=M\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{X}'=a\mathbf{XX}'$,

$$M_{\boldsymbol{\beta}}MT(\lambda)=\frac{a}{n}\operatorname{Ur}(\mathbf{I}-\mathbf{A})^2\mathbf{XX}'+\frac{\sigma^2}{n}\operatorname{Ur}\mathbf{A}^2$$

(114)

$$M_{\boldsymbol{\beta}}MV(\lambda)=\left[\frac{a}{n}\operatorname{Ur}(\mathbf{I}-\mathbf{A})^2\mathbf{XX}'+\frac{\sigma^2}{n}\operatorname{Ur}(\mathbf{I}-\mathbf{A})^2\right]\Big/\left[\frac{1}{n}\operatorname{Ur}(\mathbf{I}-\mathbf{A})\right]^2.$$

Demonstraţia urmează prin derivarea lui (114) în raport cu λ şi prin egalarea cu zero a rezultatului obţinut.

Scanned with OKEN Scanner

Capitolul XXXV

REGIUNI DE ÎNCREDERE DE DIMENSIUNE FIXATĂ PENTRU PARAMETRII REGRESIEI

§ 1. Problema I

Fie $\{y_n\}$ un şir de observaţii independente cu :

$$y_i = \boldsymbol{\beta}' \mathbf{x}^{(i)} + \varepsilon_i, \tag{1}$$

unde $\boldsymbol{\beta}'$ este un vector linie $p \times 1$ necunoscut
$\mathbf{x}^{(i)}$ un vector $p \times 1$ şi
ε_i este o variabilă avînd funcţia de repartiţie F necunoscută cu media zero şi dispersia σ^2 finită, necunoscută.

Ne propunem să găsim o regiune $\mathbf{R}$ în spaţiul Euclidian p-dimensional astfel încît

$$P\{\boldsymbol{\beta} \in \mathbf{R}\} = \alpha \tag{2}$$

şi

diametrul regiunii $\mathbf{R} \leqslant 2d$.

Deoarece nici un procedeu bazat pe o selecţie de volum fix nu satisface condiţiile puse, considerăm un procedeu secvenţial al cărui volum al selecţiei (variabilă aleatoare) N depinde de d; $N(d) \to \infty$ a.s. cînd $d \to 0$. Obţinem o regiune de încredere elipsoidală $\mathbf{R}_N$ ce satisface

i) diametrul maxim al regiunii $\mathbf{R}_N \leqslant 2d$

ii) $\lim\limits_{d \to 0} P\{\boldsymbol{\beta} \in \mathbf{R}_N\} = \alpha$.

O altă regiune de încredere $\mathbf{R}_N$, de această dată o sferă, satisface (i) şi

$$\lim_{d \to 0} P\{\boldsymbol{\beta} \in \mathbf{R}_N\} \geqslant \alpha. \tag{ii)'}$$

1.1. Soluţia

Fie $\mathbf{Y}_n' = (y_1, \ldots, y_n)$ şi $\mathbf{X}_n = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})$ o matrice $p \times n$, $p \leqslant n$ unde presupunem că $\mathbf{X}_p$ este de rang maxim (aceasta se poate realiza considerînd procesul de selecţie pînă se obţin p $x^{(i)}$ independenţi; acest volum al selecţiei va fi notat prin n_0). Estimăm parametrii necunoscuţi $\boldsymbol{\beta}$ şi σ^2 prin

$$\hat{\boldsymbol{\beta}}(n) = (\mathbf{X}_n \mathbf{X}_n')^{-1} \mathbf{X}_n \mathbf{Y}_n \tag{3}$$

Scanned with OKEN Scanner

şi respectiv

$$\hat{\sigma}^2(n) = n^{-1}\mathbf{Y}_n'\{\mathbf{I}_n - \mathbf{X}_n'(\mathbf{X}_n\mathbf{X}_n')^{-1}\mathbf{X}_n\}\mathbf{Y}_n \tag{4}$$

unde $\mathbf{I}_p$ este matricea identitate $p \times p$.

În condiţia (C) a § 1.2., găsim regiunea $\mathbf{R}_n$ cu proprietăţile i) şi ii).

Dat d şi α şi pentru un şir fixat de vectorii $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ aranjaţi astfel încît $\mathbf{X}_p$ este nesingulară, fie $\{a_n\}$ un şir de constante pozitive ce converg către numărul a ce satisface

$$P\{\chi^2(p) \leqslant a^2\} = \alpha. \tag{5}$$

Atunci acest şir $\{a_n\}$ determină un membru al clasei $\mathscr{C}$ a procedeelor secvenţiale după cum urmează.

(I) Pornim luînd $n_0 \geqslant p$ observaţii $y_1, \ldots, y_{n_0}$. Considerăm o nouă observaţie, oprind în concordanţă cu variabila de oprire N definită de

$$N = \textit{cel mai mic } k \geqslant n_0 \textit{ astfel încît } \{\hat{\sigma}^2(k) + k^{-1}\} \leqslant d^2k/a_k^2\lambda_k, \tag{6}$$

unde λ_k este rădăcina caracteristică maximă a matricii $n(\mathbf{x}_n\,\mathbf{x}_n')^{-1}$.

(II) Cînd procesul de luare a observaţiilor este oprit în $N = n$, construim regiunea $\mathbf{R}_n$ definită de

$$\mathbf{R}_n = [\mathbf{Z} : n^{-1}\{\mathbf{Z} - \hat{\boldsymbol{\beta}}(n)\}'(\mathbf{X}_n\,\mathbf{X}_n')\{\mathbf{Z} - \hat{\boldsymbol{\beta}}(n)\} \leqslant d^2/\lambda_n]. \tag{7}$$

PROPOZIŢIA 1. *Diametrul maxim al regiunii* $\mathbf{R}_n$ *este* $\leqslant 2d$.

Demonstraţie. Observăm că (7) defineşte cu probabilitatea unu interiorul şi frontiera unui elipsoid în spaţiul p-dimensional. Pentru orice elipsoid definit prin ecuaţia : $\mathbf{a}'\mathbf{D}^{-1}\mathbf{a} = 1$, diametrul maxim este egal cu $2\mu^{1/2}$, unde μ este rădăcina caracteristică maximă a matricii pozitiv definite $\mathbf{D}$. Deci diametrul maxim al elipsoidului definit prin ecuaţia

$$(n^{-1}\lambda_n/d^2)\{\mathbf{Z} - \hat{\boldsymbol{\beta}}(n)\}'(\mathbf{X}_n\,\mathbf{X}_n')\{\mathbf{Z} - \hat{\boldsymbol{\beta}}(n)\} = 1,$$

este $(2d/\lambda_n^{1/2}) = 2d$.

Albert (1966) reconsideră problema într-o generalitate mai mare. În timp ce pentru regiunile sferice el consideră de asemenea condiţia (C) şi

$$\lim_{n\to\infty} \max_{1\leqslant i\leqslant n} (\mathbf{x}^{(i)})'\mathbf{x}^{(i)} = 0 \tag{8}$$

obţine însă regiunile elipsoidale în condiţii uşor mai slabe :

$$\text{i) } \mathrm{Ur}(\mathbf{X}_n\,\mathbf{X}_n')^{-1} \to 0 \quad \text{cînd } n \to \infty$$

$$\text{ii) } \max_{1\leqslant i\leqslant n} (\mathbf{x}^{(i)\prime}\mathbf{x}^{(i)})\mathrm{Ur}(\mathbf{X}_n\,\mathbf{X}_n')^{-1} \to 0 \quad \text{cînd } n \to \infty$$

$$\text{iii) } \lim \sup_{n\to\infty} [\lambda_{\max}(\mathbf{X}_n\,\mathbf{X}_n')/\lambda_{\min}(\mathbf{X}_n\,\mathbf{X}_n')] < \infty \tag{9}$$

$$\text{iv) } \lim_{\varepsilon\to 0} \sup_n \left| \left\{ \sum_{i=1}^{n(1+\varepsilon)} (\mathbf{x}^{(i)\prime}\mathbf{x}^{(i)}) \Big/ \sum_{i=1}^{n} \mathbf{x}^{(i)\prime}\mathbf{x}^{(i)} \right\} - 1 \right| = 0,$$

Scanned with OKEN Scanner

v) Pentru un $c \geqslant 1$, $\lambda_{c(n+2)} - \lambda_{c(n+1)} \geqslant \lambda_{c(n+1)} - \lambda_{cn}$

vi) $\lambda_{n-1} \leqslant \lambda_n \to \infty$ şi $\lambda_n / \lambda_{n+1} \to 1$; $\lambda_n = \lambda_{\min}(\mathbf{X}_n \mathbf{X}_n')$.

Albert (1966) arată că, exceptînd (9.v) şi (9.vi) condiţiile sale sînt mult mai slabe decît (C) şi (8). Fără un exemplu nu este clar că condiţiile mai slabe (9) acoperă o clasă de probleme mai largă decît condiţiile (C) şi (8).

Srivastava (1971) dă un exemplu care arată că o clasă extinsă de probleme nu cere nici (C), (8), nici (9).

Exemplul 1. Considerăm problema regresiei polinomiale. Pentru uşurinţă considerăm cazul

$$y_t = \alpha + \beta t + \varepsilon_t$$

$t = 1, 2, \ldots$; unde ε_t sînt i.i.r. cu media 0 şi dispersia σ^2.
Avem

$$\mathbf{X}_n = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 2 & \ldots & n \end{pmatrix}$$

Deci

$$\mathbf{X}_n \mathbf{X}_n' = \begin{pmatrix} n & n(n+1)/2 \\ n(n+1)/2 & n(n+1)(2n+1)/6 \end{pmatrix}$$

şi $n^{-1}(\mathbf{X}_n \mathbf{X}_n') \nrightarrow a$ matrice pozitiv definită cînd $n \to \infty$. În consecinţă nu sînt satisfăcute condiţiile (C) şi (8). Pentru a arăta că nu sînt satisfăcute nici condiţiile (9), găsim rădăcinile caracteristice ale matricei $\mathbf{X}_n \mathbf{X}_n'$. Cele două rădăcini caracteristice ale lui $\mathbf{X}_n \mathbf{X}_n'$ sînt date de

$$\lambda = \{n(2n^2 + 3n + 7) \pm n[(2n^2 + 3n + 7)^2 - 12(n+1)(n-1)]^{1/2}\}/12$$

Deci

$$\lambda_{\max}(\mathbf{X}_n \mathbf{X}_n') / \lambda_{\min}(\mathbf{X}_n \mathbf{X}_n') \to \infty.$$

$\mathbf{X}_n$ poate fi scrisă ca

$$\mathbf{X}_n = \mathbf{T}_n \mathbf{L}_n \tag{10}$$

unde $\mathbf{T}_n$ este o matrice triunghiulară $p \times p$ cu elementele diagonale pozitive (deci unice) şi $\mathbf{L}_n$ este o matrice semiortogonală $p \times n$; $\mathbf{L}_n \mathbf{L}_n' = \mathbf{I}_p$ şi $\mathbf{I}_p$ este matricea identitate $p \times p$.
Deci

$$\mathbf{T}_n^{-1} \mathbf{X}_n = \mathbf{L}_n = (l_n^{(1)}, \ldots, l_n^{(n)}) = (l_{ij}(n)). \tag{11}$$

Fie

$$\lambda_n = \lambda_{\min}(\mathbf{X}_n \mathbf{X}_n'). \tag{12}$$

Scanned with OKEN Scanner

Srivastava (1971) arată că rezultatele au loc de asemenea în următoarele condiţii mai slabe :

(C*)

i) $\lim\limits_{n\to\infty} \max\limits_{1\leqslant i\leqslant n} \mathbf{l}_n^{(i)\prime}\mathbf{l}_n^{(i)} = 0,$

ii) $\lim\limits_{c\to 0} \sup\limits_{n} \left\| \sum\limits_{i=n+1}^{n^*} \mathbf{l}_{n^*}^{(i)}\mathbf{l}_{n^*}^{(i)\prime} \right\| = 0,$

unde $n^* = [n(1+c)]$ este cel mai mic întreg mai mare decît

$$n(1+c) \text{ şi } \quad \|\mathbf{B}\| = [\lambda_{\max}(\mathbf{B})]^{1/2}$$

iii) (9) (v).

PROPOZIŢIA 2. (Srivastava (1971)). *Condiţiile* (C*) *sînt mai slabe decît condiţiile* (9).

Demonstraţie. Din (11), obţinem :

$$\mathbf{l}_n^{(i)\prime}\mathbf{l}_n^{(i)} = \lambda_{\max}\mathbf{l}_n^{(i)}\mathbf{l}_n^{(i)\prime} = \lambda_{\max}\mathbf{T}_n^{-1}\mathbf{x}^{(i)}\mathbf{x}^{(i)\prime}\mathbf{T}_n^{-1} \leqslant$$

$$\leqslant [\lambda_{\max}\mathbf{B}_n^{-1}][\lambda_{\max}\mathbf{x}^{(i)}\mathbf{x}^{(i)\prime}] \leqslant \mathrm{Ur}(\mathbf{B}_n^{-1})(\mathbf{x}^{(i)\prime}\mathbf{x}^{(i)}),$$

unde $\mathbf{B}_n = \mathbf{X}_n\mathbf{X}_n' = \mathbf{T}_n\mathbf{T}_n'$. În consecinţă (9.ii) implică C*(i). În continuare, arătăm că (9.iii) şi (9.iv) implică C*(ii).

$$\left\| \sum_{n+1}^{n^*} \mathbf{l}_{n^*}^{(i)}\mathbf{l}_{n^*}^{(i)\prime} \right\| \leqslant \sum_{n+1}^{n^*} \|\mathbf{l}_{n^*}^{(i)}\mathbf{l}_{n^*}^{(i)}\| = \sum_{n+1}^{n^*} \mathbf{x}^{(i)\prime}\mathbf{B}_{n^*}^{-1}\mathbf{x}^{(i)} \leqslant$$

$$\leqslant \lambda_{\max}(\mathbf{B}_{n^*}^{-1}) \sum_{n+1}^{n^*} \mathbf{x}^{(i)\prime}\mathbf{x}^{(i)} = \left[\sum_{n+1}^{n^*} \mathbf{x}^{(i)\prime}\mathbf{x}^{(i)}\right] \Big/ \lambda_{\min}(\mathbf{B}_{n^*}) \leqslant$$

$$\leqslant k\left[\left(\sum_{n+1}^{n^*} \mathbf{x}^{(i)\prime}\mathbf{x}^{(i)}\right) \Big/ \mathrm{Ur}\mathbf{B}_{n^*}\right] \leqslant k\left[\left(\sum_{n+1}^{n^*} \mathbf{x}^{(i)\prime}\mathbf{x}^{(i)}\right) \Big/ \mathrm{Ur}\mathbf{B}_n\right].$$

Aceasta completează demonstraţia că (9) implică C*.

Acum vom verifica că condiţia C* este satisfăcută de problema regresiei polinomiale. Cu $\mathbf{X}_n\mathbf{X}_n' = \mathbf{T}_n\mathbf{T}_n'$ ca mai înainte, avem

$$\mathbf{T}_n^{-1}\mathbf{X}_n = \begin{pmatrix} n^{-1/2} & , \ldots , & n^{-1/2} \\ 0(n^{-1/2}) & , \ldots , & 0(n^{-1/2}) \end{pmatrix},$$

şi C*(i) este satisfăcută. De asemenea :

$$\left\| \sum_{i=1}^{n^*} \mathbf{l}_{n^*}^{(i)}\mathbf{l}_{n^*}^{(i)\prime} \right\|^2 \leqslant \sum_{i=n+1}^{n^*} \mathbf{l}_{n^*}^{(i)\prime}\mathbf{l}_{n^*}^{(i)} \leqslant [nc] \max_{1\leqslant i\leqslant n^*} \mathbf{l}_{n^*}^{(i)\prime}\mathbf{l}_{n^*}^{(i)} =$$

$$= [nc]0(n^{*-1}) \to 0 \text{ cînd } c \to \infty,$$

uniform în n. Pentru a arăta convexitatea condiţiei C*(iii) procedăm după cum urmează.

Scanned with OKEN Scanner

Fie

$$\lambda_n^* = 12\,\lambda_n,\ b = (2n^2+3n+7),\ \varepsilon = 12(n+1)(n-1)/b^2$$

Atunci

$$\lambda_n^* = nb[1+(1-\varepsilon)^{1/2}] = nb\left[\frac{1}{2}\varepsilon+\frac{1}{8}\varepsilon^2-\frac{1}{16}\varepsilon^3+\dots\right] =$$

$$= \frac{6n(n^2-1)}{b}+\frac{3}{2}\cdot\frac{n(n^2-1)^2}{b^2}-\frac{3}{4}\cdot\frac{n(n^2-1)^3}{b^5}+\dots$$

$$= u_1+u_2-u_3+\dots \quad (\textit{notație}).$$

Găsim

$$u_1''(x) = \frac{6(81\,x^3+113x^2+322x-21)}{b^3} > 0$$

și

$$u_2''(x) = \frac{3}{2}\,\frac{8x^7-48\,x^6-442x^5+210x^4+1000x^3-225x^2+760x+42}{b^5} > 0$$

pentru toți $x > 15$

Deci

$$\lambda_n^{*\prime\prime} = \frac{6(84n^3+113n^2+322n-21}{b^3}+\frac{3}{2}\,\frac{(8n^7-48n^6+\dots)}{b^5}-$$

$$-\,0(n^{-5})+0(n^{-7})-\dots$$

În consecință există un n_0 astfel încît pentru $n \geqslant n_0$, $\lambda_n^{*\prime\prime} \geqslant 0$ ceea ce dovedește convexitatea lui λ_n.

Pot fi construite și alte exemple care să satisfacă condițiile (C*) și să nu satisfacă (C), (8) și (9). Un alt exemplu se obține luînd matricea

$$\mathbf{X}_n = \begin{pmatrix} 0, & 1, & 0, & 2, & 0, & 3, & 0, & 4, & \dots \\ 1, & 0, & 2^{3/2}, & 0, & 3^{3/2} & 0, & 4^{3/2}, & 0, & \dots \end{pmatrix}$$

Estimatorii de C.M.M.P. pentru $\boldsymbol{\beta}$ și σ^2 sînt

$$\hat{\boldsymbol{\beta}}(n) = \mathbf{T}_n^{-1}\mathbf{L}_n\mathbf{y}_n = (\mathbf{X}_n\mathbf{X}_n')^{-1}\mathbf{X}_n\mathbf{y}_n \tag{3'}$$

și respectiv

$$\hat{\sigma}^2(n) = n^{-1}\mathbf{y}_n'[\mathbf{I}_n - \mathbf{L}_n'\mathbf{L}_n]\mathbf{y}_n = n^{-1}\mathbf{y}_n'[\mathbf{I}_n - \mathbf{X}_n'(\mathbf{X}_n\mathbf{X}_n')^{-1}\mathbf{X}_n]\mathbf{y}_n. \tag{4'}$$

Srivastava (1971) obține regiunile $\mathbf{R}_n$ și $\mathbf{R}_n'$ ce satisfac (C) și (8), după cum urmează:

(I') Pornește luînd $n_0 \geqslant p$ observații $y_1, \dots, y_{n_0}$. Se ia o nouă observație, oprindu-se în concordanță cu variabila de stopare N definită prin

$$N = \textit{cel mai mic } k \geqslant n_0 \textit{ astfel încît } (\hat{\sigma}^2(k)+k^{-1}) \leqslant d^2\lambda_k/a_k^2, \tag{6'}$$

unde λ_n este cea mai mică rădăcină caracteristică a matricei $\mathbf{X}_n\mathbf{X}_n'$, a^2 este definit ca în (5) și $a_n^2 \to a^2$.

(II′) Cînd se oprește în $N=n$, se construiește regiunea $\mathbf{R}_n$ definită de

$$\mathbf{R}_n = \{\mathbf{Z} : (\mathbf{Z} - \hat{\boldsymbol{\beta}}(n))'(\mathbf{X}_n\mathbf{X}_n')(\mathbf{Z} - \hat{\boldsymbol{\beta}}(n)) \leqslant d^2\lambda_n\}. \tag{7′}$$

PROPOZIȚIA 1′. *Diametrul maxim al regiunii* $\mathbf{R}_n$ este $\leqslant 2d$.

Pentru a obține proprietățile asimptotice ale procedeelor din clasa $\mathscr{C}$, avem nevoie de unele rezultate asimptotice date în

1.2. Teoria asimptotică pentru valori mari ale lui n

Repartiția asimptotică a statisticii $\hat{\boldsymbol{\beta}}(n)$ este obținută din următoarele rezultate

TEOREMA 1. *Dacă* $Z_1, Z_2, \ldots$ *sînt variabile* i.i.r *cu media* 0 *și dispersia* 1 *și dacă* b_{ni} $(i = 1, 2, \ldots, n;\ n = 1,2, \ldots)$ *sînt constante fixate cu*

$$\sum_{i=1}^{n} b_{ni}^2 = 1 \qquad (n = 1, 2, \ldots) \tag{13}$$

atunci dacă

$$\max_{1\leqslant i\leqslant n} |b_{ni}| \to 0, \quad \text{cînd } n \to \infty,$$

$$\lim_{n\to\infty} \mathscr{R}\left(\sum_{i=1}^{n} b_{ni}Z_i\right) = N(0, 1) \tag{14}$$

COROLARUL 1. *Fie* Σ *o matrice* $p \times p$ *pozitiv definită. Dacă are loc* (1.8) *și condiția generalizată a lui* Noether (*dată de* Gleser (1965))

$$\lim_{n\to\infty} n^{-1}(\mathbf{X}_n\mathbf{X}_n') = \boldsymbol{\Sigma}, \tag{C}$$

atunci

$$\lim_{n\to\infty} \mathscr{R}\,[(\mathbf{X}_n\mathbf{X}_n')^{1/2}\{\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}\}] = N(\mathbf{0}, \mathbf{I}_p)$$

TEOREMA 2. *Fie* $Z_1, Z_2, \ldots$ *variabile* i.i.r. *cu media* 0 *și dispersia* 1; *fie* $b_1, b_2, \ldots$ *un șir de numere astfel încît*

$$n^{-1}(b_1^2 + \ldots + b_n^2) \to 1 \text{ cînd } n \to \infty \tag{15}$$

fie $N(t)$ *o variabilă aleatoare cu valori pozitive întregi astfel încît* $N/t \to 1$ *în probabilitatea cînd* $t \to \infty$. *Atunci*

$$\lim_{\to\infty} \mathscr{R}\left(N^{-1/2}\sum_{i=1}^{N} b_iZ_i\right) = N(0, 1)$$

COROLARUL 2. *Sub condiția* (C)

$$\lim_{t\to\infty} \mathscr{R}\,(N^{-1/2}\boldsymbol{\Sigma}^{-1/2}\mathbf{X}_N\mathbf{Z}_N) = N(\mathbf{0}, \mathbf{I}_p),$$

Scanned with OKEN Scanner

unde elementele vectorului $\mathbf{Z}_N$ *sînt* i.i.r. *cu media* 0 *şi dispersia* 1 *şi* $N/t \to 1$ *în probabilitate.*

COROLARUL 3. În condiţia (C),

$$\lim_{t\to\infty} \mathscr{R}[\sigma^{-1}(\mathbf{X}_N\mathbf{X}_N')^{1/2}\{\hat{\boldsymbol{\beta}}(N) - \boldsymbol{\beta}\}] = N(\mathbf{0}, \mathbf{I}_p),$$

cînd $N/t \to 1$ *în probabilitate.*

TEOREMA 3.

$$\lim_{n\to\infty} \hat{\sigma}^2(n) = \sigma^2.$$

1.3. Proprietăţile asimptotice ale clasei $\mathscr{C}$

TEOREMA 4 (Srivastava (1967)). *Dacă este satisfăcută presupunerea* (C) *şi* $0 < \sigma^2 < \infty$, *atunci*

$$\lim_{d\to 0} d^2N/a^2\sigma^2\lambda = 1 \text{ a.s.} \tag{16}$$

$$\lim_{d\to 0} P\{\boldsymbol{\beta} \in \mathbf{R}_N\} = \alpha, \tag{17}$$

$$\lim_{d\to 0} d^2M(N)/a^2\sigma^2\lambda = 1 \tag{18}$$

Demonstraţie. Cu

$$V_n = \{\hat{\sigma}^2(n) + n^{-1}\}/\sigma^2, \tag{19}$$

$$f(n) = na^2\lambda/a_n^2\lambda_n,\ t = a^2\lambda\sigma^2/d^2, \tag{20}$$

(6) se scrie

$N = N(t) =$ *cel mai mic* $k \geqslant n_0$ *astfel încît* $V_k \leqslant f(k)/t$.

După teorema XXVII.21,

$$1 = \lim_{t\to\infty} \frac{f(N)}{t} = \lim_{d\to 0} \frac{d^2N}{\lambda_N a^2\sigma^2} = \lim_{d\to 0} \frac{d^2N}{\lambda a^2\sigma^2} \text{ a.s.} \tag{21}$$

ceea ce dovedeşte (16). Acum

$$P\{\hat{\boldsymbol{\beta}} \in \mathbf{R}_N\} = P\{[\hat{\boldsymbol{\beta}}(N) - \boldsymbol{\beta}](\mathbf{X}_N\mathbf{X}_N')[\hat{\boldsymbol{\beta}}(N) - \boldsymbol{\beta}]/\sigma^2 \leqslant d^2N/\lambda_N\sigma^2\}.$$

După (21) $d^2N/\lambda_N\sigma^2 \to a^2$ şi $N/t \to 1$ cu probabilitatea unu cînd $t \to \infty$. De asemenea după corolarul 2, pentru $t \to \infty$, avem:

$$\mathscr{R}[(\mathbf{X}_N\mathbf{X}_N')^{1/2}\{\hat{\boldsymbol{\beta}}(N) - \boldsymbol{\beta}\}] \to N(\mathbf{0}, \mathbf{I}_p).$$

Deci

$$\lim_{t\to\infty} P\{\boldsymbol{\beta} \in \mathbf{R}_N\} = \alpha$$

Din teorema 3, $\hat{\sigma}^2(n) \to \sigma^2$ a.s. În consecinţă din teorema XXVII.23 urmează (18).

Scanned with OKEN Scanner

Observații 1. Dacă cererea ii) a problemei este modificată la

ii)' $$\lim_{d\to 0} P\{\boldsymbol{\beta} \in \mathbf{R}_n'\} \geqslant \alpha$$

atunci un membru al clasei $\mathscr{C}$ definită în § 1.2, cu regiunea $\mathbf{R}_n$ modificată la

$$\mathbf{R}_n' = [\mathbf{z} : \{\hat{\boldsymbol{\beta}}(n) - \mathbf{Z}\}' \{\hat{\boldsymbol{\beta}}(n) - \mathbf{Z}\} \leqslant d^2],$$

va cere ii)'. De asemenea (16) și (18) sînt satisfăcute.

Pentru a arăta că ii)' este satisfăcută, procedăm după cum urmează :

$$P\{\boldsymbol{\beta} \in \mathbf{R}_n'\} = P\{[\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}]' [\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}] \leqslant d^2\} =$$

$$= P\{\mathbf{c}'[\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}] [\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}]' \mathbf{c} \leqslant d^2 \text{ pentru toți } \mathbf{c} : \mathbf{c}'\mathbf{c} = 1\} =$$

$$= P\left\{\frac{n^{-1}\mathbf{c}'[\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}] [\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}]'\mathbf{c}}{\mathbf{c}'(\mathbf{X}_n\mathbf{X}_n')^{-1}\mathbf{c}} \leqslant \frac{n^{-1}d^2}{\mathbf{c}'(\mathbf{X}_n\mathbf{X}_n')^{-1}\mathbf{c}} \text{ pentru toți } \mathbf{c} : \mathbf{c}'\mathbf{c} = 1\right\} \geqslant$$

$$\geqslant P\left\{\frac{n^{-1}\mathbf{c}'[\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}] [\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}]'\mathbf{c}}{\mathbf{c}'(\mathbf{X}_n\mathbf{X}_n')^{-1}\mathbf{c}} \leqslant \frac{d^2}{\lambda_n} \text{ pentru toți } \mathbf{c} : \mathbf{c}'\mathbf{c} = 1\right\} =$$

$$= P\left\{\sup_{\mathbf{c}'\mathbf{c}=1} \frac{n^{-1}\mathbf{c}'[\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}] [\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}]'\mathbf{c}}{\mathbf{c}'(\mathbf{X}_n\mathbf{X}_n')^{-1}\mathbf{c}} \leqslant \frac{d^2}{\lambda_n}\right\} =$$

$$= P\{n^{-1}[\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}]'(\mathbf{X}_n\mathbf{X}_n')[\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}] \leqslant d^2/\lambda_n\} =$$

$$= P\{[\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}]'(\mathbf{X}_n\mathbf{X}_n')[\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}] \leqslant d^2 n/\lambda_n\}.$$

Deci

$$\lim_{d\to 0} P\{\boldsymbol{\beta} \in \mathbf{R}_N'\} \geqslant \lim_{d\to 0} P\{[\hat{\boldsymbol{\beta}}(N) - \boldsymbol{\beta}]'(\mathbf{X}_n\mathbf{X}_n')[\hat{\boldsymbol{\beta}}(N) - \boldsymbol{\beta}] \leqslant d^2 N/\lambda_N\}.$$

$\mathbf{R}_N$ este un elipsoid în timp ce $\mathbf{R}_N'$ este o sferă.

2. Ca și în XXVII.9, N va fi definit ca cel mai mic întreg, impar, par, etc $\geqslant n_0$ astfel că (6) are loc.

TEOREMA 4' (Srivastava (1971)). *Dacă este satisfăcută condiția* (C*), *atunci*

a) $N(d) < \infty$ a.s. ;

b) $\lim_{d\to 0} N(d) = \infty$ a.s. ;

c) $\lim_{d\to 0} d^2\lambda_{N(d)}/\sigma^2 a^2_{N(d)} = 1$ a.s. ;

d) $\lim_{d\to 0} N(d)/q(a^2\sigma^2/d^2) = 1$ a.s. ;

unde $q(t) = \max\{n : \lambda_n < t\}$.

e) $\lim_{d\to 0} P\{\boldsymbol{\beta} \in \mathbf{R}_n\} = \alpha$

Scanned with OKEN Scanner

În demonstrație se folosește

LEMA 1. *In condiția* C*(i),

a) $\lambda_n \equiv \lambda_{min}(\mathbf{X}_n\mathbf{X}_n') \to \infty$

b) $\lambda_n/\lambda_{n+1} \to 1$, *cînd* $n \to \infty$

Demonstrație. a) Fie

$$\mathbf{B}_n = \mathbf{X}_n\mathbf{X}_n', \quad \lambda_n = \lambda_{min}(\mathbf{X}_n\mathbf{X}_n'). \tag{22}$$

În condiția că $\mathbf{X}_p = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(p)})$ este de rang maxim, $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(p)}$ sînt vectori constanți nenuli cunoscuți liniar independenți și independenți de n. Deci

$$\mathrm{Ur} \sum_{i=1}^{p} \mathbf{l}_n^{(i)}\mathbf{l}_n^{(i)'} = \sum_{i=1}^{p} \mathbf{l}_n^{(i)'}\mathbf{l}_n^{(i)} = \mathrm{Ur}\left[\left(\sum_{i=1}^{p} \mathbf{x}^{(i)}\mathbf{x}^{(i)'}\right)\mathbf{T}_n^{-1'}\mathbf{T}_n^{-1}\right] =$$

$$= \mathrm{Ur}(\mathbf{B}_p\mathbf{B}_n^{-1}) \geqslant \lambda_{\min}(\mathbf{B}_p)\mathrm{Ur}\mathbf{B}_n^{-1}.$$

Deoarece $\sum_{i=1}^{p} \mathbf{l}_n^{(i)'}\mathbf{l}_n^{(i)} \to 0$, cînd $n \to \infty$ și deoarece $\lambda_{\min}(\mathbf{B}_p)$ este independentă de n, $\mathrm{Ur}\,\mathbf{B}_n^{-1} \to 0$. Deci $\lambda_n \to \infty$.

b) Deoarece

$$\mathbf{X}_{n+1}\mathbf{X}_{n+1}' = \mathbf{X}_n\mathbf{X}_n' + \mathbf{x}^{(n+1)}\mathbf{x}^{(n+1)'}, \tag{23}$$

obținem :

$$\lambda_{n+1} \geqslant \lambda_n \quad \text{și} \quad \lambda_{n+1}/\lambda_n \geqslant 1. \tag{24}$$

Luînd inversa în ambii membrii ai relației (23) și reamintind că $\mathbf{B}_n = (\mathbf{X}_n\mathbf{X}_n')$, $\lambda_{\max}(\mathbf{B}) = \|\mathbf{B}\|^2$, obținem

$$\mathbf{B}_{n+1}^{-1} = \mathbf{B}_n^{-1/2}[\mathbf{I} - \mathbf{B}_n^{-1/2}\mathbf{x}^{(n+1)}\mathbf{x}^{(n+1)'}\mathbf{B}_n^{-1/2}]\mathbf{B}_n^{-1/2}.$$

Deci

$$\lambda_{n+1}^{-1} = \|\mathbf{B}_{n+1}^{-1}\|^2 \geqslant \|\mathbf{B}_n^{-1}\|^2[1 - \mathbf{x}^{(n+1)'}\mathbf{B}_n^{-1}\mathbf{x}^{(n+1)}] =$$

$$= \lambda_n^{-1}[1 - \mathbf{l}_{n+1}^{(n+1)'}\mathbf{T}_{n+1}'\mathbf{B}_n^{-1}\mathbf{T}_{n+1}\mathbf{l}_{n+1}^{(n+1)}] =$$

$$= \lambda_n^{-1}[1 - \mathbf{l}_{n+1}^{(n+1)'}(\mathbf{I} - \mathbf{T}_{n+1}^{-1}\mathbf{x}^{(n+1)}\mathbf{x}^{(n+1)'}\mathbf{T}_{n+1}^{-1'})^{-1}\mathbf{l}_{n+1}^{(n+1)}] =$$

$$= \lambda_n^{-1}[1 - \mathbf{l}_{n+1}^{(n+1)'}(\mathbf{I} - \mathbf{l}_{n+1}^{(n+1)}\mathbf{l}_{n+1}^{(n+1)'})^{-1}\mathbf{l}_{n+1}^{(n+1)}] \geqslant$$

$$\geqslant \lambda_n^{-1}[1 - \mathbf{l}_{n+1}^{(n+1)'}\mathbf{l}_{n+1}^{(n+1)}/(1 - \mathbf{l}_{n+1}^{(n+1)'}\mathbf{l}_{n+1}^{(n+1)}))]$$

Deci, din C*(i) urmează ca

$$\lim_{n\to\infty} \lambda_{n+1}/\lambda_n \leqslant 1 \tag{25}$$

Combinînd (24) și (25), obținem $\lambda_{n+1}/\lambda_n \to 1$ cînd $n \to \infty$.

Scanned with OKEN Scanner

§ 2. Problema 2

Considerăm un şir $\{\mathbf{X}_n\}$ $(n = 1,2,\ldots)$ de vectori p-dimensionali i.i.r. cu funcţia de repartiţie F necunoscută (posibil) cu vectorul medie $\boldsymbol{\theta}$ finit, însă necunoscut şi covarianţele σ_{ij} $(i,j = 1,2,\ldots,p)$. Adică

$$M(\mathbf{X}) = \boldsymbol{\theta} \tag{26}$$

$$\mathscr{C}(\mathbf{X}) = \boldsymbol{\Sigma} \tag{27}$$

unde $\boldsymbol{\Sigma}$ este o matrice $p \times p$ pozitiv definită de elemente σ_{ij}.

Ne propunem să găsim o regiune $\mathbf{R}$ în spaţiul Euclidian p-dimensional astfel încît

$$P\{\boldsymbol{\theta} \in \mathbf{R}\} = \alpha$$

şi astfel ca diametrul maxim al regiunii $\mathbf{R} \leqslant 2d$.

Deoarece nu există nici un procedeu bazat pe o selecţie de volum fix care să satisfacă cererile formulate, vom considera un procedeu secvential al cărui volum al selecţiei N va depinde de d; $N(d) \to \infty$ a.s. cînd $d \to 0$. Obţinem o regiune de încredere $\mathbf{R}_N$ sferică ce satisface

i) diametrul maxim al regiunii $\mathbf{R}_N \leqslant 2d$,

ii) $\lim\limits_{d \to 0} P\{\boldsymbol{\theta} \in \mathbf{R}_N\} \geqslant \alpha$.

2.1. Soluţia

Fie

$$n\overline{\mathbf{X}}_n = \sum_{i=1}^{n} \mathbf{X}_i, \tag{28}$$

şi

$$n\mathbf{S}_n = \sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}}_n)(\mathbf{X}_i - \overline{\mathbf{X}}_n)'. \tag{29}$$

Fie $a_1, a_2, \ldots$ un şir de constante pozitive ce satisface (5). Atunci acest şir $\{a_n\}$ determină un membru al clasei $\mathscr{C}^*$ de procedee secvenţiale definite după cum urmează:

Pornind luînd $n_0 > p$ observaţii $\mathbf{X}_1, \ldots, \mathbf{X}_{n_0}$. Luăm o nouă observaţie şi oprim în acord cu variabila de stopare N definită de

$$N = \textit{cel mai mic } k \geqslant n_0 \textit{ astfel încît } (\hat{\lambda}_k + k^{-1}) \leqslant (d^2 k)/a_k^2, \tag{30}$$

unde $\hat{\lambda}_n$ este rădăcina caracteristică maximă a lui $\mathbf{S}_n$. Rădăcina caracteristică maximă a lui $\boldsymbol{\Sigma}$ va fi notată prin λ.

Scanned with OKEN Scanner

ii) Cînd selecţia este oprită în $N = n$, construim regiunea $\mathbf{R}_n$ definită de :

$$\mathbf{R}_n = \{\mathbf{Z} : (\overline{\mathbf{X}}_n - \mathbf{Z})'(\overline{\mathbf{X}}_n - \mathbf{Z}) \leqslant d^2\}. \tag{31}$$

Atunci procedeele din clasa $\mathscr{C}^*$ sînt asimptotic *eficiente* cînd $d \to 0$ şi satisfac i) şi ii). Adică are loc

TEOREMA 5 (Srivastava (1967)). *Dacă* $0 < \lambda < \infty$,

$$\lim_{d\to 0} (d^2 N)/a^2\lambda = 1 \quad \text{a.s.} \tag{32}$$

$$\lim_{d\to 0} P\{\boldsymbol{\theta} \in \mathbf{R}_N\} \geqslant \alpha \tag{33}$$

şi

$$\lim_{d\to 0} (d^2\, MN)/a^2\lambda = 1 \tag{34}$$

Demonstraţie. Cu

$$y_n = (\hat{\lambda}_n + n^{-1})\lambda \tag{35}$$

$$f(n) = na^2/a_n^2 \tag{36}$$

$$t = a^2\lambda/d^2 \tag{37}$$

(30) poate fi scris ca

$$N = N(t) = \textit{cel mai mic } k \geqslant n_0 \textit{ astfel încît } y_k \leqslant f(k)/t \tag{38}$$

Deci (32) urmează din teorema XXVII.21

Folosind faptul, că

$$\sup_{\mathbf{a}'\mathbf{a}=1} \mathbf{a}'\,\boldsymbol{\Sigma}\mathbf{a} = \lambda \tag{39}$$

avem :

$$P\{\boldsymbol{\theta} \in \mathbf{R}_N\} = P\{(\overline{\mathbf{X}}_N - \boldsymbol{\theta})'(\overline{\mathbf{X}}_N - \boldsymbol{\theta}) \leqslant d^2\} =$$

$$= P\{\mathbf{a}'(\overline{\mathbf{X}}_n - \boldsymbol{\theta})(\overline{\mathbf{X}}_n - \boldsymbol{\theta})'\mathbf{a} \leqslant d^2 \textit{ pentru toţi } \mathbf{a} \textit{ astfel încît } \mathbf{a}'\mathbf{a} = 1\} =$$

$$= P\left\{\frac{N\mathbf{a}'(\overline{\mathbf{X}}_N - \boldsymbol{\theta})(\overline{\mathbf{X}}_N - \boldsymbol{\theta})'\mathbf{a}}{\lambda} \leqslant \frac{d^2N}{\lambda} \textit{ pentru toţi } \mathbf{a} : \mathbf{a}'\mathbf{a}=1\right\} \geqslant$$

$$\geqslant P\left\{\frac{N\mathbf{a}'(\overline{\mathbf{X}}_N - \boldsymbol{\theta})(\overline{\mathbf{X}}_N - \boldsymbol{\theta})'\mathbf{a}}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}} \leqslant \frac{d^2N}{\lambda} \textit{ pentru toţi } \mathbf{a} : \mathbf{a}'\mathbf{a}=1\right\} =$$

$$= P\left\{\sup_{\mathbf{a}'\mathbf{a}=1} \frac{N\mathbf{a}'(\overline{\mathbf{X}}_N - \boldsymbol{\theta})(\overline{\mathbf{X}}_N - \boldsymbol{\theta})'\mathbf{a}}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}} \leqslant \frac{d^2N}{\lambda}\right\} =$$

$$= P\{N(\overline{\mathbf{X}}_N - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{X}}_N - \boldsymbol{\theta}) \leqslant (d^2N/\lambda)\}. \tag{40}$$

Scanned with OKEN Scanner

După (32), $d^2N/\lambda \to a^2$ și $N/t \to 1$ a.s. cînd $t \to \infty$: urmează că pentru $t \to \infty$,

$$N^{1/2}\boldsymbol{\Sigma}^{-1/2}(\bar{\mathbf{X}}_N - \boldsymbol{\theta}) \sim N(\mathbf{0}, \mathbf{I}_p) \tag{41}$$

Deci

$$\lim_{t\to\infty} P\{\boldsymbol{\theta} \in \mathbf{R}_N\} \geqslant \alpha. \tag{42}$$

(34) urmează din teorema XXVII.23 și din următoarele 4 leme.

LEMA 2. *Dacă* $\hat{\lambda}_n$ *și* λ *sînt rădăcinile caracteristice maxime pentru* S_n *și respectiv* Σ, *atunci*

$$\lim_{n\to\infty} \hat{\lambda}_n = \lambda \quad \text{a.s.}$$

LEMA 3. *Fie* $\mathbf{A}$ *și* $\mathbf{B}$ *două matrice simetrice cu* $\mathbf{B}$ *ne-negativ definită. Fie* $\alpha_1 \leqslant \ldots \leqslant \alpha_p$ *rădăcinile caracteristice ordonate ale matricii* $\mathbf{A} + \mathbf{B}$ *și* $\beta_1 \leqslant \ldots \leqslant \beta_p$ *ale matricii* $\mathbf{B}$. *Atunci*

$$\alpha_i \geqslant \beta_i \qquad (i = 1,2,\ldots p).$$

LEMA 4. *Dacă* $g(n) = (n-1)/n$ $(n > p)$ *și dacă* y_n *este definită de* (35), *atunci*

$$y_n \geqslant g(n)y_{n-1}$$

LEMA 5. $\lim_{t\to\infty} \sup \{M(NY_N)/M(N)\} \leqslant 1$

Srivastava (1971) consideră situația

$$M\mathbf{Y}_n = \boldsymbol{\beta}\mathbf{X}_n$$

$$\mathscr{C}(\mathbf{y}^{(i)}) = \boldsymbol{\Sigma}, \quad i = 1,2,\ldots, n \tag{43}$$

unde $\mathbf{Y}_n = (\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)})$ este matricea $k \times n$ a observațiilor cu vectorii coloană repartizați independent,

$\boldsymbol{\beta}$ este matricea $k \times p$ a parametrilor necunoscuți și

$\mathbf{X}_n = (x^{(1)}, \ldots, x^{(n)})$ este matricea $p \times n$, $p \leqslant n - k$ a coeficienților constanți.

Specificarea (43) poate fi schimbată la cazul unidimensional. Fie $\mathbf{a}$ un vector k-dimensional nenul. Atunci (43) poate fi schimbată la cazul unidimensional după cum urmează :

$$M\mathbf{a}'\mathbf{Y}_n = \mathbf{a}'\boldsymbol{\beta}\mathbf{X}_n$$

$$\mathscr{C}(\mathbf{a}'\mathbf{y}^{(i)}) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}. \tag{44}$$

Astfel dacă avem o regiune de încredere pentru $\boldsymbol{\beta}$, putem obține de asemenea una și pentru $\mathbf{a}'\boldsymbol{\beta}$. Cerem ca

i) regiunea de încredere $\mathbf{R}_N$ pentru $\boldsymbol{\beta}$ să fie astfel încît diametrul maxim al regiunii pentru $\mathbf{a}'\boldsymbol{\beta} \leqslant 2d$ pentru fiecare vector k-dimensional $\mathbf{a}$ nenul și

ii) $\lim_{d\to 0} P(\boldsymbol{\beta} \in \mathbf{R}_N) \geqslant \alpha$

Scanned with OKEN Scanner

Dacă este satisfăcută condiția C*, aceasta se poate atinge secvențial. Estimatorii de C.M.M.P. pentru $\boldsymbol{\beta}$ și $\boldsymbol{\Sigma}$ sînt

$$\hat{\boldsymbol{\beta}}'(n) = \mathbf{T}_n^{-1'}\mathbf{L}_n\mathbf{Y}_n' = (\mathbf{T}_n\mathbf{T}_n')^{-1}\mathbf{X}_n\mathbf{Y}_n' = (\mathbf{X}_n\mathbf{X}_n')^{-1}\mathbf{X}_n\mathbf{Y}_n' \tag{45}$$

și respectiv

$$\hat{\boldsymbol{\Sigma}}(n) = n^{-1}\mathbf{Y}_n[\mathbf{I}_n - \mathbf{L}_n'\mathbf{L}_n]\mathbf{Y}_n'. \tag{46}$$

Fie C_t o variabilă aleatoare repartizată în acord cu rădăcina caracteristică maximă a unei matrici Wishart $t \times t$ cu media $l\mathbf{I}_t$ unde $t = \min(k,p)$ și $l = \max(k,p)$. Fie $\{a_n^*\}$ orice șir de numere pozitive ce converg către a^* și satisfac

$$P\{c_t \leqslant a^{*2}\} = \alpha \tag{47}$$

Fie

$$\mu_n = \lambda_{\max}\hat{\boldsymbol{\Sigma}}((n)) \quad \text{și} \quad \mu = \lambda_{\max}(\boldsymbol{\Sigma}) \tag{48}$$

Sub C*(i), $\mu_n \to \mu$ a.s. Avem din nou pașii (1′) și (II′) cu $\hat{\sigma}^2(n)$ și a_n^2 înlocuiți prin μ_n și respectiv a_n^{*2}. Regiunile de încredere $\mathbf{R}_n$ și $\mathbf{R}_n'$ $\mathbf{R}_n \subset \mathbf{R}_n'$ sînt date de

$$\mathbf{R}_n = \{\mathbf{Z}: \lambda_{\max}[(\mathbf{Z} - \hat{\boldsymbol{\beta}}(n))'(\mathbf{X}_n\mathbf{X}_n')(\mathbf{Z} - \hat{\boldsymbol{\beta}}(n))] \leqslant d^2\lambda_n\}$$

și respectiv

$$\mathbf{R}_n' = \{\mathbf{Z}: \lambda_{\max}[(\mathbf{Z} - \hat{\boldsymbol{\beta}}(n))'(\mathbf{Z} - \hat{\boldsymbol{\beta}}(n))] \leqslant d^2\}$$

TEOREMA 6. (Srivastava (1971))

$$\lim_{d\to 0} P\{\boldsymbol{\beta} \in \mathbf{R}_n'\} \geqslant \lim_{d\to 0} P\{\boldsymbol{\beta} \in \mathbf{R}_n\} \geqslant \alpha.$$

De asemenea sînt adevărate proprietățile din teorema 4′ cu σ^2 și a^2 înlocuite prin u și respectiv a^{*2}

Scanned with OKEN Scanner

Capitolul XXXVI

REGRESIE ROBUSTĂ. ESTIMAȚIE

§ 1. Introducere

Considerăm problema clasică a C.M.M.P. : p parametrii necunoscuți $\theta_1, \theta_2, \ldots, \theta_p$ sînt estimați cu ajutorul a n observații $Y_1, Y_2, \ldots, Y_n$ de care sînt legate prin

$$Y_i = \sum_{j=1}^{p} X_{ij}\,\theta_j + e_i, \tag{1}$$

unde X_{ij} sînt coeficienți cunoscuți și e_i sînt erori (variabile aleatoare independente cu repartiții aproximativ identice).

Rezolvarea clasică a problemei constă din minimizarea expresiei

$$SS = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} X_{ij}\,\theta_j \right)^2 = \min \tag{2}$$

sau echivalent, prin rezolvarea sistemului de p ecuații, obținute derivînd (2)

$$\sum_{i=1}^{n} \left(Y_i - \sum_{k=1}^{p} X_{ik}\,\theta_k \right) X_{ij} = 0, \quad j = 1, 2, \ldots, p \tag{3}$$

(X.6)

Justificarea originală a acestei metode este oarecum circulară : „estimațiile de C.M.M.P. sînt optime dacă erorile sînt i.i.r. normal; pe de altă parte, Gauss presupune că erorile sînt normale deoarece atunci media de selecție, care în general este acceptată ca o estimație bună, tinde să fie optimă în cazuri simple (Gauss (1821))".

În cazul regresiei, neomogenitatea dispersiei variabilelor e_i și faptul că erorile au repartiții cu cozi lungi au efecte care aproape nu se pot distinge, amîndouă stricînd eficiența estimațiilor. Astfel sînt necesare alternative robuste pentru metoda C.M.M.P. În baza teoremei Gauss-Markov (teorema X.7) aceste alternative nu pot fi liniare în observații. Pentru calcule de regresie pe calculator sînt necesare procedee mecanice rapide pentru înlăturarea influenței erorilor grosolane, a ponderărilor greșite (dispersii neomogene) sau a repartiției cu cozi lungi pentru erori.

Trebuie să subliniem totuși că aceasta nu implică cercetarea atentă a rezidurilor

$$r_i = Y_i - \sum_j X_{ij}\,\theta_j. \tag{4}$$

Scanned with OKEN Scanner

Există două moduri pentru obținerea estimațiilor robuste. Primul constă în înlocuirea funcției (2) care trebuie minimizată printr-o expresie care este mai puțin sensibilă la valorile extreme ale rezidurilor.

Al doilea mod este bazat pe observația că după (3) parametrii θ_j sînt estimați astfel încît vectorul rezidual **r** și vectorii coloană $\mathbf{X}_j$ au corelațiile empirice 0 (sau mai precis momentele mixte, deoarece nu sînt scăzute din ele mediile).

Avem trei tipuri de estimatori competitori: M, L și R (prezentați pe scurt în XI. §8).

Reluăm și dezvoltăm considerațiile care au condus la acești estimatori.

Într-o repartiție descrisă de densitatea de repartiție $f(x)$ introducem un parametru de locație (de alunecare) θ necunoscut, obținînd densitatea $f(x-\theta)$. Dacă $X_1, X_2, \ldots, X_n$ reprezintă o selecție asupra variabilei avînd această repartiție, o metodă obișnuită de estimare a parametrului θ este metoda V.M. Logaritmul funcției de verosimilitate este:

$$\ln P(\theta) = \sum_{i=1}^{n} \ln f(x_i - \theta) = -\sum_{i=1}^{n} \rho(x_i - \theta),$$

unde $\rho(x) = -\ln f(x)$.

Dacă $\ln P(\theta)$ este o funcție derivabilă avem:

$$\frac{d[\ln P(\theta)]}{d\theta} = -\sum_{i=1}^{n} \frac{f'(x_i - \theta)}{f(x_i - \theta)} = \sum_{i=1}^{n} \psi(x_i - \theta),$$

unde $\rho'(x) = \psi(x)$. Soluția ecuației

$$\sum_{i=1}^{n} \psi(x_i - \theta) = 0$$

ce minimizează $P(\theta)$ poartă numele de estimatorul de verosimilitate maximă pentru θ și este notat prin $\hat{\theta}$.

Trei exemple tipice ale acestui proces sînt date de următoarele repartiții

1. *Normală*: $\rho(x) = x^2/2 + C$, $\psi(x) = x$, $\sum_{i=1}^{n} (x_i - \theta) = 0 \Rightarrow \hat{\theta} = \bar{x}$

2. *Dublu-exponențială*: $\rho(x) = |x| + C$, $\psi(x) = -1$, $x < 0$ și $\psi(x) = 1$, $x > 0$, $\sum_{i=1}^{n} \psi(x_i - \theta) = 0 \Rightarrow \hat{\theta}$ = mediana de selecție.

3. *Cauchy*: $\rho(x) = \ln(1 + x^2) + C$, $\psi(x) = 2x/(1 + x^2)$, $\sum_{i=1}^{n} \psi(x - \theta) = 0$

este rezolvată prin metode iterative.

Este interesant de notat că funcțiile ψ din exemplele 2 și 3 sînt mărginite și că cea din exemplul 3 chiar se apropie asimptotic de zero. Notăm că soluțiile în 2 și 3 nu sînt influențate mult de datele nesemnificative. Pe de altă parte, estimatorul de C.M.M.P. din 1 este puternic influențat de valorile extreme. Estimatorul de C.M.M.P. $\bar{X}$ nu este bun în situațiile în care repartiția sub cercetare are cozi lungi, de exemplu,

Scanned with OKEN Scanner

în cazul repartiției Cauchy. Estimația robustă, cere pentru estimatorii care sînt destul de eficienți (în jur de 0,90 — la 0,95) dacă repatiția este normală, să fie de asemeni foarte eficienți chiar dacă repartiția are cozi lungi. Adeseori cantitatea de eficiență pierdută sub presupunerile normale, de 0,05, este un fel de premium ce se plătește pentru protecția ce se dă în cazul în care repartiția actuală are cozi mai lungi decît cea normală.

Din motive teoretice (actual minimizarea disperisiei asimptotice maxime a estimatorilor asociați cu o anumită clasă de repartiții), Huber (1964), (1968), (1973) propune să se folosească ca estimator robust estimatorul de verosimilitate maximă a parametrului de poziție asociat cu o densitate care este aproape normală în mijloc însă aproape dublu exponențială în cozi. În particular, funcția ρ a lui Huber (exceptînd o constantă aditivă) este

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & , |x| \leqslant k \\ k|x| - \frac{1}{2}k^2, & |x| > k \end{cases} \tag{5}$$

astfel că funcția ψ este

$$\psi(x) = \begin{cases} -k, & x < -k \\ x, & |x| \leqslant k \\ k, & x > k. \end{cases}$$

Cu această funcție ψ, ecuația

$$\sum_{i=1}^{n} \psi(x_i - \theta) = 0$$

trebuie să fie rezolvată prin metode iteartive. Un estimator de acest tip se notează prin $\hat{\theta}$ și poartă numele de *estimator M*, *M* pentru verosimilitatea maximă.

Merită să fie evidențiat un alt aspect al estimatorului *M*. Presupunem că am găsit, pentru o selecție particulară $x_1, x_2, \ldots, x_n$, o soluție $\hat{\theta}$ și că am înlocuit această selecție printr-una în care abaterile de la $\hat{\theta}$, au fost triplate. Noua soluție $\hat{\theta}$, obținută folosind această nouă selecție modificată, nu va fi necesar aceeași; adică estimatorul nu va fi invariant de scară. Pentru a obține un estimator invariant, vom rezolva ecuația

$$\sum_{i=1}^{n} \psi[(x_i - \theta)/d] = 0,$$

unde d este o estimație robustă a scării. O statistică d folosită în această situație este

$$d = \text{mediana } |x_i - \text{mediana } (x_i)|/(0{,}6745).$$

Se folosește divizorul 0,6745 deoarece atunci $d \simeq \sigma$ dacă n este mare și dacă selecția actuală provine dintr-o populație normală.

Scanned with OKEN Scanner

Această schemă particulară de alegere a lui d sugerează valorile corespunzătoare ale constantei k astfel încît eficiența lui $\hat{\theta}$ să fie mare dacă repartiția este normală. În situația normală dorim ca cele mai multe observații să satisfacă inegalitatea :

$$|(x_i - \theta)/d| \leqslant k$$

deoarece atunci

$$\psi[(x_i - \theta)/d] = (x_i - \theta)/d.$$

Alte funcții ψ (remarcate de Hogg (1979)) care sînt folosite mai des sînt

1. *Hampel* (a, b, c) (1972)

$$\psi(x) = (\text{sign}x) \begin{cases} |x| & , \quad 0 \leqslant |x| < a \\ a & , \quad a \leqslant |x| < b \\ a\dfrac{c - |x|}{c - b}, & b \leqslant |x| < c \\ 0 & , \quad c \leqslant |x| \end{cases}$$

2. *Andrews* (1972)

$$\psi(x) = \begin{cases} \sin(x/k), & |x| \leqslant k\pi \\ 0 \quad , & |x| > k\pi, \end{cases}$$

cu $k = 1,5$ sau 2,0. Dacă scala este cunoscută, $k = 1,339$ cere un premium de 0,05.

3. *Tukey* (1972)

$$\psi(x) = \begin{cases} x[1 - (x/k)^2]^2, & |x| \leqslant k \\ 0 \quad , & |x| > k \end{cases}$$

cu $k = 5,0$ sau 6,0. Dacă scala este cunoscută, $k = 4,685$ implică un premium de 5%.

Deoarece funcțiile ρ asociate cu aceste trei funcții ψ nu sînt convexe, vor fi anumite probleme de convergență în procedeele iterative cu toate că această situație nu este generală.

Presupunem că printr-o metodă numerică iterativă am găsit soluția $\hat{\theta}$ a ecuației noastre ; adică fie $\hat{\theta}$ estimatorul M astfel că $\sum \psi[(X_i - \hat{\theta})/d] = 0$, pentru orice funcție ψ mărginită care este impară.

Se aproximează această ultimă ecuație înlocuind membrul stîng prin primii doi termeni ai dezvoltării în serie Taylor, în jurul valorii adevărate a parametrului θ. În această aproximație, cînd $d = \sigma$.

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{\sigma\left[\sum \psi\left(\dfrac{X_i - \theta}{\sigma}\right)\right] \Big/ \sqrt{n}}{\left[\sum \psi'\left(\dfrac{X_i - \theta}{\sigma}\right)\right] \Big/ \sqrt{n}}$$

are la limită o repartiţie normală cu media zero şi dispersia

$$\frac{\sigma^2 M\left[\psi^2\left(\frac{X-\theta}{\sigma}\right)\right]}{\left\{M\left[\psi'\left(\frac{X-\theta}{\sigma}\right)\right]\right\}^2}.$$

Deoarece nu cunoaştem densitatea, trebuie să aproximăm valorile medii şi σ ce apar în această dispersie asimptotică. O aproximaţie a dispersiei statisticii $\sqrt{n}(\hat{\theta}-\theta)$ este

$$s_1^2 = \frac{d^2\left\{(1/n)\sum_{i=1}^{n}\psi^2[(x_i-\hat{\theta})/d]\right\}}{\left\{(1/n)\sum_{i=1}^{n}\psi'[(x_i-\hat{\theta})/d]\right\}^2}.$$

Adică, pentru valori mari ale lui n,

$$\sqrt{n}(\hat{\theta}-\theta)/s_1$$

are asimptotic o repartiţie $N(0, 1)$. Huber (1970) arată că repartiţia statisticii $(n-1)^{1/2}(\hat{\theta}-\theta)/s$, poate fi aproximată asimptotic, printr-o repartiţie t, cu $n-1$ grade de libertate (sau chiar una cu mai puţin de $n-1$ grade de libertate) şi aceasta ar fi o aproximaţie chiar mai bună decît cea normală. Gross (1976), bazat pe această idee cercetează intervale de încredere.

Dacă ne reîntoarcem la modelul (1), înlocuim (2) prin

$$\sum_{i=1}^{n}\rho(Y_i-\sum_j X_{ij}\theta_j)=\min, \tag{6}$$

unde ρ este dat de (5).

Dacă derivăm (6), obţinem analog lui (3) ecuaţia

$$\sum_{i=1}^{n}\psi(Y_i-\sum_k X_{ik}\theta_k)X_{ij}=0,\ j=1,\ldots,p, \tag{7}$$

care este echivalentă cu (6) dacă ρ este convexă.

Huber (1973) se ocupă de estimaţiile definite de (6) şi (7). Yohai, Maronna (1979) studiază comportarea asimptotică a estimatorilor M.

Estimarea R este o metodă neparametrică rezultată din atribuirea de ranguri elementelor unei selecţii ce provin dintr-o repartiţie de tip continuu. Estimarea R se extinde la regresie. În acest scop considerăm modelul liniar (1) şi modificăm metoda C.M.M.P. din (2) înlocuind un factor din $\left(Y_i-\sum_{j=1}^{p}X_{ij}\theta_j\right)^2$ prin rangul lui $r_i=\left(Y_i-\sum_{j=1}^{p}X_{ij}\theta_i\right)1\leqslant i\leqslant n$; fie R_i acest rang. Rangul R_i este o funcţie de θ. Deci trebuie să minimizăm

$$\sum_{i=1}^{n}\left(Y_i-\sum_{j=1}^{p}X_{ij}\theta_j\right)R_i.$$

Scanned with OKEN Scanner

Aceasta la rîndul ei poate fi generalizată înlocuind rangurile $1, 2, \ldots, n$ prin scorurile :

$$a_n(1) \leqslant a_n(2) \leqslant \ldots \leqslant a_n(n).$$

Astfel în această generalizare, trebuie să minimizăm

$$\sum_i \left(Y_i - \sum_{j=1}^{p} X_{ij}\, \theta_j \right) a_n(R_i) = \min. \tag{8}$$

Dacă derivăm (8), care este o funcţie liniară convexă de $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ obţinem (Jaeckel (1972)) că această minimizare este echivalentă cu a rezolva sistemul de p ecuaţii

$$\sum_i X_{ij} a(R_i) = 0, \qquad j = 1, \ldots, p \tag{9}$$

ce trebuiesc rezolvate aproximativ datorită discontinuităţilor în $a_n(\cdot)$ şi R_i.

Scorurile bune sînt cele date de

$$a(i) = \varphi[i/(n+1)],$$

unde

$$\varphi(t) = -f'[F^{-1}(t)]/f[F^{-1}(t)].$$

Exemple de astfel de scoruri sînt :

1. f normală produce $\varphi(t) = \Phi^{-1}(t)$, $0 < t < 1$, adică dă scorurile normale

2. f dublu exponenţială produce $\varphi(t) = -1$, $0 < t < 1/2$ şi $\varphi(t) = 1$, $1/2 < t < 1$, ce dau scorurile mediană şi

3. f logistică produce $\varphi(t) = 2t - 1$, $0 < t < 1$, ce dă scorurile Wilcoxon.

Ecuaţiile (9) conduc la problema de minim ;

$$\sum_j \left| \sum_i X_{ij}\, a_n(R_i) \right| = \min. \tag{10}$$

Ultima variantă a fost cercetată de Jurečková (1971) şi echivalenţa asimptotică dintre (8) şi (10) a fost demonstrată de Jaeckel (1972). În loc de a corela $a_n(R_i)$ cu X_{ij} ca în (9), putem folosi rangurile pentru $\mathbf{X}_j$.

Putem ataca problema generală a regresiei, considerînd un parametru în timp şi fixînd valorile celorlalţi. Totuşi nu este uşor de stabilit existenţa unui punct fix şi chiar dacă există unul, procesul iterativ pentru el nu este necesar convergent.

Mai mult, toate aceste estimaţii ale regresiei sînt versiuni cu un pas : se porneşte cu o estimaţie preliminară bună θ^* şi se aplică metoda Newton relaţiei (7).

Fie $\hat{\Delta}_M$ estimatorul M şi $\hat{\Delta}_R$ estimatorul R al parametrului $\Delta = (\Delta_1, \ldots, \Delta_p)$ din modelul regresiei liniare din §4 presupunerea 1°. În anumite condiţii asupra matricei (c_{ij}), asupra funcţiei de repartiţie F a variabilelor independente şi asupra funcţiilor ce generează estimatorii respectivi, Jurečková (1977) obţine repartiţia asimptotică a statisticii

Scanned with OKEN Scanner

$\hat{\Delta}_M - \hat{\Delta}_R$ pentru p fixat și $N \to \infty$. Rezultatul obținut de Jurečková (1977) are mai multe consecințe care la rîndul lor prezintă interes ; printre altele arată că la orice estimator M corespunde un estimator R, estimatori care sînt asimptotic echivalenți. Un caz interesant apare cînd $\hat{\Delta}_M$ este estimatorul de verosimilitate maximă corespunzător unei repartiții G și $\hat{\Delta}_R$ este estimația de rang care este asimptotic eficientă pentru G în rolul repartiției variabilelor independente. Dacă G este normală, G este normală, atunci $\hat{\Delta}_M$ și $\hat{\Delta}_R$ sînt asimptotic echivalenți dacă și numai dacă F este de asemenea normală. Jurečková (1977) arată că cele două estimații coincid dacă G este dublu exponențială (atunci estimațiile reprezintă o generalizare a medianei).

Hogg (1979) remarcă că generalizarea estimatorilor L la situațiile regresionale nu este atît de clară ca în cazul estimatorilor M. Cum folosirea funcției $\rho(x) = |x|$, dă mediana ca estimator, aceasta poate fi modificată pentru a da alte percentile. Astfel funcția ρ

$$\rho(x) = -(1-p)\,x, \qquad x < 0$$

dă percentila de ordin 100 p în cazul unei singure selecții și acestea pot fi construite în problemele de regresie.

Sievers (1979) propune estimații robuste pentru parametrii din modelul liniar, estimații bazate pe statistici de rang ponderate. Metoda este bazată pe minimizarea funcției de dispersie definită ca media lui Gini ponderată. Se arată liniaritatea asimptotică și se obține repartiția asimptotică a estimatorului astfel obținut.

§ 2. Estimații M

2.1. Metoda C.M.M.P. cînd erorile nu sînt normale

În cele ce urmează urmăm Eicker (1963), (1967), Huber (1973).

Presupunem că problema de regresie este înserată într-un șir similar de probleme cu matricile

$$\mathbf{X}^{(N)} = (x_{ij}^N),\ 1 \leqslant i \leqslant n_N, \quad 1 \leqslant j \leqslant p_N, \qquad N \to \infty.$$

Presupunem că erorile U_i sînt independente și identic repartizate, avînd o repartiție nenormală fixată cu media 0 și dispersia $\sigma^2 < \infty$. Atunci estimația de C.M.M.P. pentru $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ este

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \qquad (\text{X}.7) \tag{11}$$

Problema care se pune este de a găsi condiții necesare și suficiente astfel încît toate estimațiile de forma

$$\hat{\alpha} = \sum a_j \hat{\theta}_j$$

să fie asimptotic normale.

Scanned with OKEN Scanner

Putem alege sistemul de coordonate astfel încît el să fie ortogonal :

$$\mathbf{X}'\mathbf{X} = \mathbf{I}. \tag{12}$$

De asemenea putem standardiza a_j astfel încît $\sum a_j^2 = 1$. Atunci

$$\hat{\alpha} = \sum_i s_i Y_i \tag{13}$$

cu

$$s_i = \sum_j x_{ij} a_j \tag{14}$$

şi

$$\sum s_i^2 = 1, \tag{15}$$

deci

$$D^2(\hat{\alpha}) = \sigma^2. \tag{16}$$

LEMA 1. *Estimaţia $\hat{\alpha}$ este asimptotic normală dacă*

$$\max_i |s_i| \to 0 \text{ cînd } N \to \infty.$$

Demonstraţie. Dacă $\max_i |s_i| \not\to 0$, atunci repartiţia limită a lui $\hat{\alpha}$ (dacă există) poate fi descompusă în două componente dintre care una nu este normală ; deci ea însăşi nu poate fi normală. Pe de altă parte, dacă $\max |s_i| \to 0$, atunci are loc condiţia lui Lindeberg, care în acest caz este

$$\frac{1}{\sigma^2} \sum_i M\{s_i^2 U_i^2 1_{(|s_i U_i| > \varepsilon\sigma)}\} \to 0. \tag{17}$$

Inegalitatea lui Schawrz dă :

$$s_i^2 = (\sum_j x_{ij} a_j)^2 \leqslant \sum_j x_{ij}^2 \sum_j a_j^2 = \sum_j x_{ij}^2$$

egalitatea avînd loc dacă $a_j \sim x_j$.
Observăm că

$$\gamma_{ii} = \sum_j x_{ij}^2 \tag{18}$$

este elementul diagonal i al matricei proiecţie

$$\Gamma = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{19}$$

(care este independentă de sistemul de coordonate particular din spaţiul parametrilor).

Deci are loc :

PROPOZIŢIA 1. *Condiţia necesară şi suficientă pentru ca toate estimaţiile $\hat{\alpha}$ de C.M.M.P. să fie asimptotic normale este*

$$\max_i \gamma_{ij} \to 0 \text{ cînd } N \to \infty. \tag{20}$$

Scanned with OKEN Scanner

Observații. i) Deoarece $\sum_i \gamma_{ii} = p$, avem

$$\max \gamma_{ii} \geqslant p/n$$

și deci (20) implică $p/n \to 0$.

ii) Valoarea ajustată a lui Y_i (estimația de C.M.M.P. a mediei lui $\mathbf{X}_i$) este

$$T_i = \sum_j x_{ij} \hat{\theta}_j = \sum_l \gamma_{il} Y_l.$$

În particular, dacă (20) nu are loc, atunci unele dintre valorile ajustate nu vor fi asimptotic normale.

iii) Rezidurile sînt

$$Y_i - T_i = (1 - \gamma_{ii}) Y_i - \sum_{l \neq i} \gamma_{il} Y_l.$$

Deci dacă γ_{ii} sînt apropiate de 1, o eroare mare în Y_i nu arată neapărat o creștere în rezidul $Y_i - T_i$, însă ea poate arăta o creștere în altă parte să spunem în rezidul $Y_m - T_m$ dacă γ_{mi} este mare!. Astfel putem spune că (20) asigură un fel de *robustețe a planului*. Robusteţea planului este optimizată luînd $\gamma_{ii} = p/n$ pentru toți i; atunci vom numi matricea $\mathbf{X}$ *balansată*. Notăm că condiția

$$\limsup p/n < 1$$

este o condiție suficientă pentru ca γ_{ii} să fie întotdeauna mărginite de 1.

În baza acestor observații, condiția (20) (și deci $p/n \to 0$) apare ca indispensabilă pentru orice teorie asimptotică generală a regresiei robuste.

2.2. Estimații *M* ale regresiei

Huber (1973) estimează parametrii necunoscuți minimizînd o expresie de forma (6). Lucrează în următorele condiții de regularitate:

1) *Condiții asupra matricei* $\mathbf{X}$. Elementele diagonale ale matricei proiecție

$$\Gamma = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

sînt presupune a fi uniform mici:

$$\max_{1 \leqslant i \leqslant n} \gamma_{ii} = \varepsilon \ll 1,$$

ordinul de precizie al micimii fiind specificat de la caz la caz. Fără a pierde din generalitate putem alege sistemul de coordonate în spațiul parametrilor astfel încît punctul parametru adevărat este $\theta^\circ = 0$, și că $\mathbf{X}'\mathbf{X}$ este matricea identitate.

2) *Condiții asupra estimației.* Se presupune că funția ρ este convexă și nu monotonă, că posedă derivate de ordin suficient de mari mărginite. În particular,

$$\psi(x) = (\mathrm{d}/\mathrm{d}x)\rho(x)$$

va fi continuă și mărginită.

Scanned with OKEN Scanner

Convexitatea lui ρ serveşte pentru garantarea echivalenţei între (6) şi (7) şi pentru unicitatea soluţiei asimptotice ; în rest nu este importantă. Derivatele de ordin superior sînt necesare pentru dezvoltări Taylor, însă existenţa lor nu este esenţială pentru ca rezultatele să aibă loc.

3) *Condiţii asupra repartiţiilor erorii.* Presupunem că erorile U_i sînt i.i.r., astfel încît

$$M(\psi(U_i)) = 0.$$

Această presupunere este necesară pentru ca media lui (6) să atingă minimul în valoarea adevărată θ°.

Presupunerea de independenţă este o restricţie serioasă ; presupunerea că erorile sînt identic repartizate simplică notaţiile şi calculele.

Cazurile $\varepsilon p^2 \to 0$ şi $\varepsilon p \to 0$.

Se obţin rezultate bune numai dacă $\varepsilon p^2 \to 0$ şi rezultate mai puţin satisfăcătoare dacă $\varepsilon p \to 0$. Aceasta implică $p^3/n \to 0$ şi respectiv $p^2/n \to 0$; astfel valorile moderate ale lui p conduc la valori mari şi nepracticabile pentru n. Ideea este de a compara zerourile a 2 funcţii aleatoare vectori $\Phi(\theta)$ şi $\Psi(\theta)$:

$$\Phi_j(\theta) = \frac{-1}{M(\psi')} \sum_i \psi(Y_i - \sum_k x_{ik}\,\theta_k)\, x_{ij}, \tag{21}$$

$$\Psi_j(\theta) = \theta_j - \frac{1}{M(\psi')} \sum_i \psi(Y_i)\, x_{ij} \tag{22}$$

Zeroul $\tilde{\theta}$ al funcţiei Ψ este

$$\tilde{\theta}_j = \frac{1}{M(\psi')} \sum_i \psi(Y_i)\, x_{ij}, \tag{23}$$

care nu este o estimaţie pură. După rezultatele din § 2.1, toate combinaţiile liniare

$$\tilde{\alpha} = \sum a_j\, \tilde{\theta}_j$$

sînt asimptotic normale dacă $\varepsilon \to 0$. Zeroul $\hat{\theta}$ al funcţiei Φ este estimaţia considerată şi vom arăta că diferenţa dintre $\hat{\theta}$ şi $\tilde{\theta}$ este mică. Fie a_j coeficienţii nedeterminaţi ce satisfac

$$\sum a_j^2 = 1$$

şi fie s_i definit de (14). Scriem

$$t_i = \sum_j x_{ij}\, \theta_j. \tag{24}$$

Deoarece $\mathbf{X'X} = \mathbf{I}$, avem

$$\| \mathbf{t} \|^2 = \sum t_i^2 = \sum \theta_j^2 = \| \theta \|^2. \tag{25}$$

Scanned with OKEN Scanner

Dezvoltăm $\sum a_j \Phi_j(\theta)$ în serie Taylor

$$\sum_j a_j \Phi_j(\theta) =$$

(26)

$$= \frac{-1}{M(\psi')} \left\{ \sum \psi(Y_i) s_i - \sum \psi'(Y_i) t_i s_i + \frac{1}{2} \sum \psi''(Y_i - \eta t_i) t_i^2 s_i \right\}$$

cu $0 < \eta < 1$. Aceasta poate fi aranjată pentru a da

$$(27) \quad \sum a_j(\Phi_j(\theta) - \Psi_j(\theta)) = \sum_{j,k} \Delta_{jk} a_j \theta_k - \frac{1}{2M(\psi')} \sum_i \psi''(Y_i - \eta t_i) t_i^2 s_i$$

unde

$$(28) \quad \Delta_{jk} = \frac{1}{M(\psi')} \sum_i [\psi'(Y_i) - M\psi'(Y_i)] x_{ij} x_{ik}.$$

Acum arătăm că expresia (27) este mică pe mulţimile de forma

$$(29) \quad \{(\theta, \mathbf{a}) \mid \| \theta \|^2 \leqslant Kp, \ \| \mathbf{a} \| = 1\}.$$

Din inegalitatea lui Schwarz

$$(30) \quad (\sum \Delta_{jk} a_j \theta_k)^2 \leqslant \sum_{j,k} \Delta_{jk}^2 \sum_j a_j^2 \sum_k \theta_k^2 = \sum_{j,k} \Delta_{jk}^2 \| \theta \|^2.$$

Avem :

$$(31) \quad M(\sum_{j,k} \Delta_{jk}^2) = \sum_{j,k} M(\Delta_{jk}^2) = \sum_{j,k,i} x_{ij}^2 x_{ik}^2 \frac{D^2(\psi')}{(M\psi')^2}$$

şi

$$(32) \quad \sum_{j,k,i} x_{ij}^2 x_{ik}^2 = \sum_i \gamma_{ii}^2 \leqslant \max \gamma_{ii} \sum_i \nu_{ii} = \varepsilon p.$$

Fie $\delta > 0$; atunci inegalitatea lui Markov arată că

$$(33) \quad K_1 = \frac{D^2(\psi')}{(M\psi')^2} \cdot \frac{1}{\delta}$$

astfel încît

$$(34) \quad P(\sum_{j,k} \Delta_{jk}^2 \geqslant K_1 \varepsilon p) \leqslant \delta.$$

Urmează, că cu o probabilitate $> 1 - \delta$,

$$(35) \quad (\sum_{j,k} \Delta_{jk} a_j \theta_k)^2 \leqslant K K_1 \varepsilon p^2$$

au loc simultan pentru toţi $(\mathbf{a}, \theta)$ din (29).

Presupunem că ψ'' este mărginită, şi anume

$$|\psi''(x)| \leqslant 2M(\psi')E \text{ pentru unii } E,$$

Scanned with OKEN Scanner

atunci

$$(36)\qquad \left|\frac{1}{2M(\psi')}\sum_i \psi''(Y_i-\eta t_i)\,t_i^2 s_i\right| \leqslant E\max|s_i|\sum t_i^2 \leqslant E\varepsilon^{1/2}\|\theta\|^2$$

(a se vedea (25) şi (18)).

Cu acestea, obţinem cu o probabilitate $>1-\delta$, (27) este mărginit în valoare absolută de

$$(37)\qquad r=((KK_1)^{1/2}+EK)(\varepsilon p^2)^{1/2}$$

şi această mărginire este uniformă pe mulţimea (29). Deoarece rezultatul are loc simultan pentru toţi $\mathbf{a}$ cu $\|\mathbf{a}\|=1$, am arătat că cu o probabilitate $>1-\delta$

$$(38)\qquad \|\Phi(\theta)-\Psi(\theta)\|\leqslant r \text{ pentru } \|\theta\|^2\leqslant Kp.$$

Cum

$$(39)\qquad M(\|\tilde{\theta}\|^2)=\frac{M(\psi^2)}{(M\psi')^2}p,$$

din inegalitatea lui Markov urmează că : $\|\tilde{\theta}\|^2\leqslant Kp/4$ cu o probabilitate mare, cu condiţia ca K să fie ales mare. Atunci

$$(40)\qquad \|\Phi(\theta)-\theta\|\leqslant\|\Phi(\theta)-\Psi(\theta)\|+\|\tilde{\theta}\|\leqslant r+\frac{1}{2}(Kp)^{1/2}$$

pe mulţimea $\|\theta\|^2\leqslant Kp$.

Dacă $\varepsilon p\to 0$, r poate fi făcut mai mic decît $\frac{1}{2}(Kp)^{1/2}$, astfel încît (40) implică

$$(41)\qquad \|\Phi(\theta)-\theta\|\leqslant\|\theta\|,$$

pe sfera $\|\theta\|^2=Kp$, şi din teorema de punct fix a lui Brouwer urmează că $\Phi(\theta)$ are un zero în $\|\theta\|^2<Kp$.

Mai mult, dacă inserăm $\theta=\hat{\theta}$ în (38), obţinem că

$$(42)\qquad \|\hat{\theta}-\tilde{\theta}\|\leqslant r.$$

În particular, dacă $\varepsilon p^2\to 0$, aceasta implică

$$(43)\qquad \|\hat{\theta}-\tilde{\theta}\|\xrightarrow{P}0.$$

Dacă numai $\varepsilon p\to 0$, obţinem rezultatul mai slab

$$(44)\qquad \frac{\|\hat{\theta}-\tilde{\theta}\|}{p^{1/2}}\xrightarrow{P}0.$$

Scanned with OKEN Scanner

Observație. În baza lui (39), $\|\tilde{\boldsymbol{\theta}}\| \sim p^{1/2}$.

Fie $\hat{\alpha} = \sum a_j \hat{\theta}_j$ și $\tilde{\alpha} = \sum a_j \tilde{\theta}_j$, cu $\|\mathbf{a}\| = 1$. Reamintim că $\hat{\alpha}$ este estimație ce trebuie cercetată, în timp ce $\tilde{\alpha}$ este o sumă de variabile aleatoare independente și dacă $\varepsilon \to 0$, este asimptotic repartizată normal.

PROPOZIȚIA 2. *Dacă* $\varepsilon p^2 \to 0$, *atunci*

$$\sup_{\|\mathbf{a}\|=1} |\hat{\alpha} - \tilde{\alpha}| \xrightarrow{P} 0. \tag{45}$$

Dacă $\mathbf{a}$ *este ales aleator în raport cu o măsură invariantă pe sfera* $\|\mathbf{a}\| = 1$, *și dacă* $\varepsilon p \to 0$, *atunci*

$$\hat{\alpha} - \tilde{\alpha} \xrightarrow{P} 0. \tag{46}$$

În particular, (46) *implică că* $\hat{\alpha}$ *este asimptotic normal.*

Demonstrație. (45) urmează din (43); (46) urmează din (44) și din faptul că media lui $|\hat{\alpha} - \alpha|^2$ pe sfera unitate $\|\mathbf{a}\| = 1$ este $\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\|^2/p$.

Observație. Presupunerea că valoarea adevărată a punctului parametru este $\boldsymbol{\theta}^0 = \mathbf{0}$ a fost folosită numai în (39). Dacă $\boldsymbol{\theta}^*$ este orice estimație ce satisface

$$\|\boldsymbol{\theta}^* - \boldsymbol{\theta}^0\| = \mathbf{0}_p(p^{1/2}),$$

atunci se poate arăta în același mod ca în metoda lui Newton (pentru rezolvarea ecuației $\Phi(\boldsymbol{\theta}) = 0$, cu valoarea $\boldsymbol{\theta}^*$) că se obține estimația $\hat{\boldsymbol{\theta}}^*$ ce satisface

$$\|\hat{\boldsymbol{\theta}}^* - \tilde{\boldsymbol{\theta}}\| \xrightarrow{P} 0, \quad \|\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}\| \xrightarrow{P} 0,$$

cu condiția ca $\varepsilon p^2 \to 0$.

Huber (1973) afirmă că (46) are loc pentru orice $\mathbf{a}$ fixat sau orice alegere aleatoare a lui $\mathbf{a}$ care este independentă de observații, cu condiția ca $\varepsilon p \to 0$. Pe de altă parte, apare că (46) nu este în general adevărată dacă $\varepsilon p \not\to 0$.

Cu $p = 1$, găsim *dezvoltări formale asimptotice în serii de puteri*, dezvoltări necesare în următoarele două paragrafe.

Fie

$$\lambda(t) = M[\psi(U + t)].$$

Vom nota valorile ajustate prin

$$T_i = \sum_j x_{ij} \hat{\theta}_j,$$

unde $(\hat{\theta}_j)$ este soluția lui (7).

Ideea este de a dezvolta (7) într-o serie de puteri în jurul valorii adevărate $\boldsymbol{\theta} = \mathbf{0}$ și de a găsi dezvoltări asimptotice a cantităților interesante ordonate după puterile lui ε.

Rescriem (7) ca

$$\hat{\theta}_j = \sum_i \frac{\psi(Y_i)}{\lambda'(0)} x_{ij} + R_j(t), \tag{47}$$

Scanned with OKEN Scanner

unde

$$R_j(T) = \sum_i \frac{1}{\lambda'(0)} [\psi(Y_i - T_i) - \psi(Y_i) + \lambda'(0) T_i] x_{ij}.$$

Fie

$$V_{ki} = \frac{\psi^{(k)}(Y_i) - \lambda^{(k)}(0)}{\lambda'(0)}, \quad k \geqslant 0$$

$$\beta_k = \lambda^{(k)}(0)/\lambda'(0)$$

şi dezvoltăm termenul rămas într-o serie de puteri

$$R_j(T) = \sum_{k \geqslant 1} R_{kj}(T),$$

cu

$$R_{kj}(T) = \frac{(-1)^k}{k!} \sum_i V_{ki} T_i^k x_{ij} + \frac{(-1)^{k+1}}{(k+1)!} \beta_{k+1} \sum_i T_i^{k+1} x_{ij}.$$

Acum (47) se scrie

$$\hat{\theta}_j = \sum_i V_{0i} x_{ij} + R_j(t).$$

Pornim procedeul, punînd

$$\hat{\theta}_j^{(1)} = \sum_i V_{0i} x_{ij}$$

$$T_i^{(k)} = \sum_j \hat{\theta}_j^{(k)} x_{ij}$$

$$\hat{\theta}_j^{(k+1)} = \sum_i V_{0i} x_{ij} + R_j(T^{(k)}). \tag{48}$$

Putem privi aceste cantităţi ca dezvoltări formale în serie de puteri în V_{ki}.

Termenii de ordinul k în V sînt aceeaşi ca în $\hat{\theta}^{(k)}$, $\hat{\theta}^{(k+1)}$, ..., astfel că procedeul converge în mod formal. Obţinem aceeaşi expresie dacă înlocuim formula de recurenţă (48) prin

$$\hat{\theta}_j^{(k+1)} = \hat{\theta}_j^{(1)} + [R_{1j}(T^{(k)}) + R_{2j}(T^{(k-1)}) + \ldots + R_{kj}(T^{(1)})]_{k+1} \tag{49}$$

unde $[\ \ ]_{k+1}$ arată că numai termeni de ordin $\leqslant k+1$ în V sînt reţinuţi. Deoarece (49) conţine numai polinoame în V, calculele sînt mai uşoare. Fie U_{kj} suma tuturor termenilor de ordinul k din această dezvoltare a lui $\hat{\theta}_j$, astfel ca:

$$\tilde{\theta}_j^{(k)} = U_{1j} + U_{2j} + \ldots + U_{kj}.$$

Scanned with OKEN Scanner

Huber (1973) obține

$$U_{1j} = \sum_i V_{0i}\, x_{ij}$$

$$U_{2j} = -\sum_{i,l} V_{0i}\, V_{1l}\, \gamma_{il}\, x_{ij} + \frac{\beta_2}{2} \sum_{i,l,m} V_{0i}\, V_{0m}\, \gamma_{il}\, \gamma_{im}\, x_{ij}$$

$$U_{3j} = \sum V_{0m}\, V_{1i}\, V_{1l}\, \gamma_{il}\, \gamma_{im}\, x_{ij} - \frac{\beta_2}{2} V_{0m}\, V_{0g}\, V_{1i}\, \gamma_{il}\, \gamma_{im}\, \gamma_{ig}\, x_{ij} -$$

$$- \beta_2 \sum V_{0m}\, V_{0g}\, V_{1l}\, \gamma_{il}\, \gamma_{im}\, \gamma_{ig}\, x_{ij} + \frac{\beta_2^2}{2} \sum V_{0m}\, V_{0g}\, V_{0s}\, \gamma_{il}\, \gamma_{im}\, \gamma_{ig}\, \gamma_{is}\, x_{ij} +$$

$$+ \frac{1}{2} V_{0m}\, V_{0g}\, V_{2i}\, \gamma_{im}\, \gamma_{ig}\, x_{ij} - \frac{\beta_3}{6} \sum V_{0m}\, V_{0g}\, V_{0s}\, \gamma_{im}\, \gamma_{ig}\, \gamma_{is}\, x_{ij}.$$

În cazul simetric (cînd erorile U_i sînt repartizate simetric și ρ este simetric față de 0) coeficienții β_k cu numere pare dispar. Pentru motive de simetrie

$$M(U_{kj}) = 0$$

și avem

$$M(U_{kj}^2) = \mathbf{0}(\varepsilon^{k-1}). \tag{50}$$

2.3. Deplasarea

În § 2.2. am standardizat problema astfel încît, ideal, estimațiile $\hat{\theta}_j$ vor avea o repartiție normală nedegenerată centrată în 0 însă în cazul asimetric pot apărea *deplasări*.

Primul termen al deplasării este

$$M(\hat{\theta}_j^{(2)}) = M(U_{2j}) = K_0 \sum_i \gamma_{ii}\, x_{ij},$$

cu

$$K_0 = -M(V_0 V_1) + \frac{\beta_2}{2} M(V_0^2).$$

(am pus $V_0, V_1, \ldots$, în loc de $V_{0i}, V_{1i}, \ldots$).

Termenul deplasare corespunzător lui $\hat{\alpha} = \sum a_i \hat{\theta}_j$, cu $\sum a_j^2 = 1$ este

$$M(\hat{\alpha}^{(2)}) = K_0 \sum_i \gamma_{ii}\, s_i,$$

cu $s_i = \sum_j x_{ij}\, a_j$. Inegalitatea lui Schwarz dă:

$$|\sum_i \gamma_{ii}\, s_i| \leqslant (\sum_i \gamma_{ii}^2)^{1/2} \leqslant (\varepsilon p)^{1/2}.$$

Scanned with OKEN Scanner

Aceasta nu trebuie să fie mică cînd ε este mic şi deplasarea poate fi foarte serioasă. De exemplu, luăm cazul balansat $\gamma_{ii} = p/n$ şi presupunem că θ_1 corespund la efectul principal ce afectează egal toate observaţiile, adică $x_{i1} \equiv 1/n^{1/2}$. Atunci :

$$M(\hat{\theta}_1^{(2)}) = K_0 \frac{p}{n^{1/2}}$$

(51)

$$M(\hat{\theta}_j^{(2)}) = 0, \qquad j > 1,$$

şi deplasarea lui $\hat{\theta}_1^{(2)}$ nu tinde la zero, în afară de cazul în care $\varepsilon p = p^2/n \to 0$

Media deplasării valorii ajustate T_i este

$$M(T_i^{(2)}) = K_0 \sum_l \gamma_{il}\, \gamma_{ll}. \tag{52}$$

Inegalitatea lui Schwarz dă

$$|\sum_l \gamma_{il}\, \gamma_{ll}| \leqslant (\sum_l \gamma_{il}^2 \sum_l \gamma_{ll}^2)^{1/2} \leqslant \varepsilon p^{1/2}. \tag{53}$$

Următoarea matrice arată că această margine este asimptotic

$$\mathbf{X} = \begin{pmatrix} \mathbf{e} & 0 & \dots & 0 & 0 \\ 0 & \mathbf{e} & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & \mathbf{e} & 0 \\ 0 & 0 & \dots & 0 & \mathbf{e} \\ g & g & \dots & g & g \end{pmatrix} \tag{54}$$

unde $\mathbf{e}$ reprezintă un vector coloană cu toate cele r componente egale cu 1 şi $g = r/(n - p)$.

Adică presupunem că $n = pr + 1$, că fiecare θ_j este observat de r ori, şi că există o observaţie adiţională Y_n de $g(\theta_1 + \dots + \theta_p)$. Acest $\mathbf{X}$ nu trebuie să satisfacă $\mathbf{X}'\mathbf{X} = \mathbf{I}$, însă $\gamma_{ii} \equiv p/n$ şi

$$\sum_l \gamma_{il}\, \gamma_{ln} = \left(\frac{p}{n}\right)^2 [1 + (r(n-p))^{1/2}] \sim \frac{p}{n} p^{1/2}. \tag{55}$$

O deplasare mică în $\hat{\theta}_j$ poate să ducă la o creştere mare a deplasării în $T_n = g(\hat{\theta}_1 + \dots + \hat{\theta}_p)$.

Astfel, un rezidu izolat $Y_n - T_n$ poate avea cauza nu o eroare mare în Y_n, ci o deplasare mare în valoarea ajustată T_n !.

2.4. Termenii covarianţă

Calculele arată că în cazul simetric

$$D^2(U_{kj}) = \mathbf{0}(\varepsilon^{k-1}), \qquad k = 1, 2, 3. \tag{56}$$

Scanned with OKEN Scanner

Urmează că

$$\operatorname{cov}(\hat{\theta}_j, \hat{\theta}_k) = \operatorname{cov}(\hat{\theta}_j^{(1)}, \hat{\theta}_k^{(1)}) + \mathbf{0}(\varepsilon), \tag{57}$$

$$\operatorname{cov}(\hat{\theta}_j, \hat{\theta}_k) = \operatorname{cov}(\hat{\theta}_j^{(3)}, \hat{\theta}_k^{(3)}) + \mathbf{0}(\varepsilon^2) \tag{58}$$

și

$$\begin{aligned}\operatorname{cov}(\hat{\theta}_j^{(3)}, \hat{\theta}_k^{(3)}) = {} & \operatorname{cov}(U_{1j}, U_{1k}) + \operatorname{cov}(U_{2j}, U_{2k}) + \\ & + \operatorname{cov}(U_{1j}, U_{2k}) + \operatorname{cov}(U_{1j}, U_{3k}) + \\ & + \operatorname{cov}(U_{2j}, U_{1k}) + \operatorname{cov}(U_{3j}, U_{1k}) + \mathbf{0}(\varepsilon^2).\end{aligned}$$

Vom evalua (59) în cazul asimetric

$$\begin{aligned}\operatorname{cov}(\hat{\theta}_j^{(3)}, \hat{\theta}_k^{(3)}) = {} & M(V_0^2)\,\delta_{jk} + K_1 \sum_i \gamma_{ii}\, x_{ij}\, x_{ik} + K_2 \sum_{i,l} \gamma_{il}^2\, x_{ij}\, x_{lk} + \\ & + K_3 \sum_{i,l} \gamma_{ii}\, \gamma_{ll}\, x_{ij}\, x_{lk} + \mathbf{0}(\varepsilon^2)\end{aligned} \tag{60}$$

cu

$$K_1 = 3M(V_0^2)M(V_1^2) - 2M(V_0^2V_1) + 3M(V_0^2)M(V_0V_2) - \beta_3(M(V_0^2))^2$$

$$K_2 = 3(M(V_0V_1))^2 + \beta_2 M(V_0^3) - 8\beta_2 M(V_0^2)M(V_0V_1) + \frac{5\beta_2^2}{2}(M(V_0^2))^2$$

$$K_3 = 2(M(V_0V_1))^2 - 3\beta_2 M(V_0^2)M(V_0V_1) + \beta_3^2(M(V_0^2))^2.$$

În cazul simetric avem

$$K_2 = K_3 = 0.$$

Fie $\hat{\alpha} = \sum a_j \hat{\theta}_j$ cu $\sum a_j^2 = 1$ și $s_i = \sum x_{ij} a_j$.
Atunci

$$\begin{aligned}D^2(\hat{\alpha}) = {} & M(V_0^2) + K_1 \sum_i \gamma_{ii}\, s_i^2 + K_2 \sum_{i,l} \gamma_{il}^2\, s_i\, s_l + \\ & + K_3 \sum_i \Big(\sum_l \gamma_{ll}\, \gamma_{ll}\Big)\, s_i^2 + \mathbf{0}(\varepsilon^2).\end{aligned} \tag{61}$$

Avem

$$0 \leqslant \sum_i \gamma_{ii}\, s_i^2 \leqslant \varepsilon$$

PROPOZIȚIA 3. *Al 3-lea termen din* (61) *satisface*

$$0 \leqslant \sum_{i,l} \gamma_{il}^2\, s_i\, s_l \leqslant \varepsilon.$$

Demonstrație. Fie (t_i) vectorul propriu ce aparține celei mai mari valori proprii λ a lui (γ_{il}^2).

Atunci

$$\lambda t_i = \sum_l \gamma_{il}^2 t_l \leqslant \sum_l \gamma_{il}^2 \max_l t_l = \gamma_{i\cdot} \max_l t_l.$$

Astfel

$$\lambda \max_l t_l \leqslant \varepsilon \max_l t_l.$$

deci $\lambda \leqslant \varepsilon$, şi astfel:

$$0 \leqslant \sum_{i,l} \gamma_{i\cdot l}^2 s_i s_l \leqslant \varepsilon.,$$

Însă al 4-lea termen cauzează greutăţi. Nu se cunoaşte dacă sau nu marginea obişnuită

$$|\sum_i \sum_l \gamma_{il} \gamma_{il} s_i^2| \leqslant \varepsilon p^{1/2}$$

(care urmează din (53)) este exactă. Cu matricea (54) şi $a_j = 1/p^{1/2}$ se obţine

$$\sum_i \sum_l \gamma_{il} \gamma_{il} s_i^2 \sim \left(\frac{p}{n}\right)^2 p^{1/2} = \varepsilon^2 p^{1/2}.$$

În orice caz, în general termenul al treilea nu este de ordinul $\mathbf{0}(\varepsilon)$, şi astfel în cazul asimetric, presupunerile (57) şi (58) nu pot fi amîndouă adevărate.

2.5. Estimaţia matricii de covarianţe a lui $\hat{\boldsymbol{\theta}}$

Huber (1973) foloseşte ca estimaţie a matricii de covarianţe a lui $\hat{\boldsymbol{\theta}}$ o matrice de forma

$$\frac{\frac{1}{n-p}\sum \psi(Y_i - T_i)^2}{\left[\frac{1}{n}\sum \psi'(Y_i - T_i)\right]^2} (\mathbf{X}'\mathbf{X})^{-1} \tag{62}$$

Adeseori, numitorul $n - p$ nu este corespunzător şi $\mathbf{X}'\mathbf{X}$ va fi înlocuită printr-o matrice proporţională cu $\mathbf{W} = (W_{jk})$:

$$W_{jk} = \sum_i \psi'(Y_i - T_i) x_{ij} x_{ik}. \tag{63}$$

Aici o dezvoltare formală în serie de puteri poate da unele lămuriri euristice.

Vom presupune că $p \to \infty$ astfel că $1/n$ este neglijabil faţă de p/n.

i) *Termenul* $n^{-1} \sum \psi'(Y_i - T_i)$. Dezvoltarea lui $\psi'(Y_i - T_i)/\lambda'(0)$ este

$$\frac{\psi'(Y_i - T_i)}{\lambda'(0)} = 1 + V_{1i} - \beta_2 T_i - V_{2i} T_i + \frac{\beta_3}{2} T_i^2 + \ldots \tag{64}$$

Scanned with OKEN Scanner

Pentru deplasări, avem

$$M(T_i) = K_0 \sum_i \gamma_{ii}\gamma_{ii} + \dots \tag{65}$$

$$M(V_{2i}T_i) = M(V_0V_2)\gamma_{ii} + \dots \tag{66}$$

$$M(T_i^2) = M(V_0^2)\gamma_{ii} + \dots \tag{67}$$

Astfel, după sumarea după i

$$M\left\{\frac{1}{n}\sum \psi'(Y_i - T_i)\right\} = \lambda'(0)\left\{1 - \frac{p}{n}\left[K_4 + \frac{\sum_{i,l}\gamma_{ii}\gamma_{ii}}{p}K_5\right]\right\} + \dots \tag{68}$$

cu

$$K_4 = M(V_0V_2) - \frac{\beta_3}{2}M(Y_0^2)$$

$$K_5 = \beta_2 K_0.$$

Dispersia lui $n^{-1}\sum\psi'(Y_i - T_i)$ este neglijabilă fiind de ordinul $0(1/n)$

ii) *Termenul* $n^{-1}\sum\psi(Y_i - T_i)^2$. Dezvoltarea fiind:

$$\frac{\psi(Y_i - T_i)^2}{\lambda'(0)^2} = V_{0i}^2 + T_i^2(1 + V_{1i})^2 - 2T_iV_{0i}(1 + V_{1i}) + \tag{69}$$

$$+ T_i^2V_{0i}(\beta_2 + V_{2i}) + \dots$$

pentru deplasări, avem

$$M(T_i^2(1 + V_{1i})^2) = \gamma_{ii}M(V_0^2)(1 + MV_{1i}^2) + \dots$$

$$M(2T_iV_{0i}(1 - V_{1i})) = \gamma_{ii}(2M(V_0^2) + 2M(V_0^2V_1) +$$

$$+ 2(\sum_l \gamma_{ll}\gamma_{li})K_0M(V_0V_1) + \dots$$

$$M(T_i^2V_{0i}(\beta_2 + V_{2i})) = \gamma_{ii}M(V_0^2)M(V_0V_2) + \dots$$

Astfel obţinem:

$$M\left[\frac{1}{n}\sum\psi(Y_i - T_i)^2\right] = \tag{70}$$

$$= \lambda'(0)^2M(V_0^2)\left\{1 - \frac{p}{n}\left[1 - K_6 - \frac{\sum_{i,l}\gamma_{ll}\gamma_{li}}{p}K_7\right]\right\} + \dots$$

cu

$$K_6 = M(V_1^2) - 2\frac{M(V_0^2V_1)}{M(V_0^2)} + M(V_0V_2)$$

$$K_7 = 2\frac{(M(V_0V_1))^2}{M(V_0^2)} - \beta_2M(V_0V_1) = -2\frac{M(V_0V_1)}{M(V_0^2)}K_0.$$

Scanned with OKEN Scanner

Din nou dispersia lui $n^{-1} \sum \psi(Y_i - T_i)^2$ este neglijabilă, fiind de ordinul $0(1/n)$, astfel

$$M\left[\frac{1}{n-p} \sum \psi(Y_i - T_i)^2\right] =$$

(71)

$$= (\lambda'(0))^2 M(V_0^2) \left\{1 + \frac{p}{n}\left[K_6 + \frac{\sum_{i,l} \gamma_{il}\gamma_{il}}{p} K_7\right]\right\} + \dots$$

iii) *Matricea* **W**. Presupunem $\mathbf{X'X} = \mathbf{I}$ şi definim **W** ca în (63); obţinem

$$\frac{\mathbf{W}}{\lambda'(0)} = \mathbf{I} + \mathbf{B}, \tag{72}$$

dezvoltarea matricii **B** fiind

$$B_{jk} = \sum_i \left[V_{1i} - (\beta_2 + V_{2i})\, T_i + \frac{1}{2}\, \beta_3 T_i^2\right] x_{ij}\, x_{ik} + \dots \tag{73}$$

Sîntem interesaţi în $\mathbf{W} = \lambda'(0)(\mathbf{I} + \mathbf{B})$

$$\mathbf{W}^{-1} = \frac{1}{\lambda'(0)}(\mathbf{I} - \mathbf{B} + \mathbf{B}^2 - \dots) \tag{74}$$

$$\mathbf{W}^{-2} = \frac{1}{\lambda'(0)^2}(\mathbf{I} - 2\mathbf{B} + 3\mathbf{B}^2 - \dots).$$

Pentru deplasare obţinem

$$M(B_{jk}) = -\sum_i [K_4 \gamma_{ii} + K_5 \sum_l \gamma_{il}\gamma_{il}]\, x_{ij}\, x_{ik} + \dots$$

(75)

$$M(\sum_m B_{jm} B_{mk}) = M(V_1^2) \sum_i \gamma_{ii}\, x_{ij}\, x_{ik} + 2K_5 \sum_{i,l} \gamma_{il}^2\, x_{ij}\, x_{lj}.$$

Astfel, în particular

$$M(W_{jk}^{-1}) = \frac{1}{\lambda'(0)}\{\delta_{jk} + [K_4 + M(V_1^2)] \sum_i \gamma_{ii}\, x_{ij}\, x_{ik} + \tag{76}$$

$$+ K_5 \sum_i (\sum_l \gamma_{il}\gamma_{il})\, x_{ij}\, x_{ik} + 2K_5 \sum_{i,l} \gamma_{il}^2\, x_{ij}\, x_{lk}\} + \dots$$

$$M(W_{jk}^{-2}) = \frac{1}{\lambda'(0)^2}\{\delta_{jk} + [2K_4 + 3M(V_1^2)] \sum_i \gamma_{ii}\, x_{ij}\, x_{ik} + \tag{77}$$

$$+2K_5 \sum_i (\sum_l \gamma_{il}\gamma_{il})\, x_{ij}\, x_{ik} + 6K_5 \sum_{i,l} \gamma_{il}^2\, x_{ij}\, x_{lk}\} \dots$$

Scanned with OKEN Scanner

Huber (1973) propune pentru estimarea matricei de covarianţe a lui $\hat{\boldsymbol{\theta}}$ unul dintre estimatorii :

$$\frac{\frac{1}{n-p}\sum \psi(Y_i - T_i)^2}{\left[\frac{1}{n}\sum \psi'(Y_i - T_i)\right]^2} (\mathbf{X}'\mathbf{X})^{-1} \tag{78}$$

$$\frac{\frac{1}{n-p}\sum \psi(Y_i - T_i)^2}{\frac{1}{n}\sum \psi'(\mathbf{Y}_i - T_i)} \mathbf{W}^{-1} \tag{79}$$

$$\frac{1}{n-p}\sum \psi(Y_i - T_i)^2 \mathbf{W}^{-1}(\mathbf{X}'\mathbf{X})\mathbf{W}^{-1}. \tag{80}$$

Toate trei se reduc la expresia clasică în cazul $\psi(x) = x$. Toate trei concordă dacă $p = 1$, $x_{ij} = 1/n^{1/2}$.

Introducînd expresiile obţinute în (i), (ii), (iii) în aceste estimaţii, le comparăm cu matricea de covarianţe (60); formulele astfel obţinute nu sînt uşor de comparat. Totuşi dacă presupunem că repartiţia erorilor şi ρ sînt simetrice şi că matricea $\mathbf{X}$ este balansată ($\gamma_{ii} \equiv p/n$), avem o simplificare. Toate cele 3 estimaţii ale covarianţei conţin o deplasare de ordinul $0(p/n)$ şi este posibil să le facem nedeplasate pînă la termenii $0(p^2/n^2)$, dacă le multiplicăm prin factorul de corecţie K^2, K sau respectiv K^{-1}, unde

$$K = 1 + \frac{p}{n} M(V_1^2). \tag{81}$$

În cazul de faţă vom înlocui $M(V_1^2)$ printr-o estimaţie a sa.

2.6. Calcularea unor astfel de estimaţii

Folosim

$$\psi(x) = \max(-c, \min(c, x)), \tag{82}$$

şi estimăm simultan θ şi σ (prametru de scară). Deoarece dispersia estimatorului lui σ va fi de ordinul $\mathbf{0}(1/n) = 0(p/n)$, această modificare nu influenţează teoria asimptotică care a fost obţinută pentru σ fixat.

Propunem să rezolvăm sistemul neliniar

$$\sum_i \psi\left(\frac{(Y_i - \sum_k x_{ik}\,\theta_k}{\sigma}\right) x_{ij} = 0, \qquad j = 1, 2, \ldots, p. \tag{83}$$

$$\frac{1}{n-p}\sum_i \psi\left(\frac{Y_i - \sum_k x_{ik}\,\theta_k}{\sigma}\right)^2 = \beta \tag{84}$$

cu

$$\beta = M_\Phi \psi(U)^2$$

Sînt două metode principale pentru rezolvarea lui (83), una brută şi una mai sofisticată.

Mai întîi metoda brută. Se liniarizează (83)

$$\sum_i \sigma\psi\left(\frac{Y_i}{\sigma}\right)x_{ij} - \sum_k \sum_i \psi'\left(\frac{Y_i}{\sigma}\right) x_{ij}\, x_{ik}\, \theta_k = 0, \tag{85}$$

se înlocuieşte $\psi'(Y_i/\sigma)$ prin valoarea sa medie (sau printr-o estimaţie a sa) şi apoi se rezolvă pentru θ. Pentru a creşte acurateţea, procedeul poate fi repetat cu Y_i înlocuit prin rezidul $Y_i - T_i$.

Mai precis, procedeul poate fi descris după cum urmează :

1. Se calculează $(\mathbf{X}'\mathbf{X})^{-1}$
2. Se iau valorile de început, θ, σ
3. Fie

$$T_i = \sum_k x_{ik}\, \theta_k, \quad m = \frac{1}{n} \sum \psi'\left(\frac{Y_i - T_i)}{\sigma}\right)$$

$$Y_j = \sum_i \sigma\psi\left(\frac{Y_i - T_i}{\sigma}\right) x_{ij}, \quad q = \frac{1}{n-p} \sum \psi\left(\frac{Y_i - T_i}{\sigma}\right)^2 \sigma^2.$$

4. Punem

$$\sigma_{\text{nou}} = \left(\frac{q}{\beta}\right)^{1/2}$$

$$\Delta\theta = \frac{1}{m} (\mathbf{X}'\mathbf{X})^{-1} V$$

$$\theta_{\text{nou}} = \theta + \Delta\theta$$

5. Dacă $\| \Delta\theta \|$ este mai mică decît o constantă prestabilită, stop. În rest ne întoarcem la 2 cu noi valori pentru θ şi σ.

Metoda sofisticată. Fie $\theta^{(r)}$, $\sigma^{(r)}$ valori succesive ; atunci în (82) sînt trei clase de reziduri

clasa inferioară $I_-^{(r)} = \{i |\, Y_i - \sum x_{ik}\, \theta_k^{(r)} < - x\sigma^{(r)}\}$

clasa de mijloc $I_0^{(r)} = \{i |\, - x\sigma^{(r)} \leqslant Y_i - \sum x_{ik}\, \theta_k^{(r)} \leqslant x\sigma^{(r)}\}$

clasa superioară $I_+^{(r)} = \{i |\, Y_i - \sum x_{ik}\, \theta_k^{(r)} > x\sigma^{(r)}\}$.

Dacă, partiţia $I_-^{(r)}$, $I_0^{(r)}$, $I_+^{(r)}$ concordă cu aceea indusă de soluţia lui (83), (84), atunci această ultimă soluţie poate fi obţinută într-un singur pas. Pentru a vedea aceasta, scriem sistemul ca :

$$\sum_0 (Y_i - \sum_k x_{ik}\, \theta_k)\, x_{ij} + (\sum_+ x_{ik} - \sum_- x_{ik}) x\sigma = 0 \tag{68}$$

$$\sum_0 (Y_i - \sum_k x_{ik}\, \theta_k)^2 + (\sum_+ 1 + \sum_- 1)\, x^2\sigma^2 = (n-p)\beta\sigma^2, \tag{87}$$

unde 0, +, − arată că sumarea se face după $I_0^{(r)}$, $I_+^{(r)}$ şi respectiv $I_-^{(r)}$.

Scanned with OKEN Scanner

Procedeul de calcul este:

1. Alegem valorile iniţiale $\theta^{(0)}$, $\sigma^{(0)}$. Fie $r = 0$
2. Găsim mulţimile $I_0^{(r)}$, $I_+^{(r)}$, $I_-^{(r)}$.
3. Calculăm matricea

$$W_{jk} = \sum_0 x_{ij}\, x_{ik}, \tag{88}$$

vectorii

$$V_k = \sum_0 Y_i\, x_{ik}$$

$$R_k = \sum_+ x_{ik} - \sum_- x_{ik} \tag{89}$$

şi scalarii

$$Q = \sum_0 Y_i^2$$

$$E = \sum_+ 1 + \sum_- 1 \tag{90}$$

4. Rezolvăm

$$\sum W_{jk}\, \tau_k = V_j, \quad j = 1, 2, \ldots, p$$

$$\sum W_{jk}\, \delta_k = R_j, \quad j = 1, 2, \ldots, p \tag{91}$$

pentru τ şi pentru δ.

5. Atunci $\theta_k(\sigma) = \tau_k + x\sigma\delta_k$ satisfac (86) identic în σ; dacă inserăm această funcţie liniară în (87) şi rezolvăm pentru σ^2, obţinem:

$$(\sigma^{(r+1)})^2 = \frac{\sum_0 (Y_i - \sum_k x_{ik}\, \tau_k)^2}{(n-p)\beta + x^2 [\sum_0 (\sum_k x_{ik}\, \delta_k)^2 + E]} \tag{92}$$

sau echivalent (însă mai sensibil la erorile de rotunjire)

$$(\sigma^{(r+1)})^2 = \frac{Q - \sum_k V_k\, \tau_k}{(n-p)\beta - x^2 [\sum_k R_k\, \delta_k + E]} \tag{93}$$

6. Punem $\theta_k^{(r+1)} = \tau_k + x\sigma^{(r+1)}\delta_k$ şi găsim partiţia corespunzătoare $I_0^{(r+1)}$, $I_+^{(r+1)}$, $I_-^{(r+1)}$. Dacă aceasta concordă cu precedenta $I_0^{(r)}$, $I_+^{(r)}$, $I_-^{(r)}$, atunci $(\theta^{(r+1)}, \sigma^{(r+)})$ rezolvă (83). În caz contrar se înlocuieşte r prin $r+1$ şi ne întoarcem la 3.

Pentru a mări acurateţea, putem înlocui în (89) Y_i prin valoarea curentă a rezidului $Y_i - \sum_k x_{ik}\, \theta_k^{(r)}$. Atunci

$$\theta_k^{(r+1)} = \theta_k^{(r)} + \tau_k + x\sigma^{(r+1)}\, \delta_k.$$

Deoarece avem numai un număr finit de partiţii posibile I_0, I_+, I_-, procedeul poate fi oprit după un număr finit de paşi, sau poate fi repetat periodic.

Scanned with OKEN Scanner

Procedeul sofisticat este rapid convergent (3 la 5 iteraţii). Pentru valori mari ale lui p/n, matricea $\mathbf{W}$ devine singulară. Procedeul brut necesită între 15 şi 30 iteraţii.

În final, după § 2.5, matricea de covarianţe a lui $\hat{\boldsymbol{\theta}}$ poate fi estimată de

$$\frac{\frac{1}{n-p}\sum\psi\left(\frac{Y_i-T_i}{\sigma}\right)^2\sigma^2}{\frac{1}{n}\sum\psi'\left(\frac{Y_i-T_i}{\sigma}\right)}K\mathbf{W}^{-1} \tag{94}$$

sau de

$$\frac{\frac{1}{n-p}\sum\psi\left(\frac{Y_i-T_i}{\sigma}\right)\sigma^2}{\left[\frac{1}{n}\sum\psi'\left(\frac{Y_i-T_i}{\sigma}\right)\right]^2}K^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{95}$$

2.7. Comportarea asimptotică a estimatorilor *M* pentru modelul liniar

Yohai, Maronna (1979) consideră modelul liniar

$$\mathbf{y}_n = \mathbf{X}_n\,\boldsymbol{\theta} + \mathbf{u}_n, \tag{96}$$

unde $\mathbf{X}_n$ este o matrice dată de ordin $n \times p$,

$\boldsymbol{\theta}$ este un vector p-dimensional necunoscut al coeficienţilor de regresie,

$\mathbf{u}_n$ este un vector cu componentele $(u_1, \ldots, u_n)$, unde u_i sînt variabile aleatoare independente identic repartizate cu funcţia de repartiţie F şi

$\mathbf{y}_n = (y_{1n}, \ldots, y_{nn})$ este vectorul observaţiilor.

După cum am văzut (§ 1, § 2.2) estimatorii M ai lui Huber sînt definiţi ca soluţii ale

$$\sum_{i=1}^{n}\rho((y_{in} - \mathbf{x}'_{in}\,\hat{\boldsymbol{\theta}}) = \min, \tag{97}$$

unde funcţia ρ este aleasă convenabil şi $\mathbf{x}'_{in} \in R^p$ este linia i a matricei $\mathbf{X}_n$ cu elementele $x_{jin}(j = 1, \ldots, p;\ i = 1, \ldots, n)$. Dacă ρ este convexă cu derivata ψ, atunci (97) este echivalentă cu

$$\sum_{i=1}^{n}\psi(y_{in} - \mathbf{x}'_{in}\,\hat{\boldsymbol{\theta}}_{in})\,\mathbf{x}_{in} = \mathbf{0}. \tag{98}$$

Aceşti estimatori nu sînt invarianţi de scară, adică nu este satisfăcută

$$\hat{\boldsymbol{\theta}}(a\,y_1, \ldots, a\,y_n) = a\,\hat{\boldsymbol{\theta}}(y_1, \ldots, y_n).$$

Scanned with OKEN Scanner

Pot fi obţinuţi estimatori invarianţi ca soluţii a

$$\sum_{i=1}^{n} \psi_{\hat{\sigma}}(y_{in} - \mathbf{x}'_{in}\hat{\boldsymbol{\theta}})\,\mathbf{x}_{in} = 0, \tag{99}$$

unde

$$\psi_\sigma(x) = \psi(x/\sigma),$$

şi $\hat{\sigma}(y_1, \ldots, y_n)$ este un estimator al lui σ ce satisfce

$$\hat{\sigma}(ay_1, \ldots, ay_n) = |a|\hat{\sigma}(y_1, \ldots, y_n).$$

Considerăm condiţiile

(A1) ψ este nedescrescătoare

(A2) există numerele pozitive b, c şi d astfel încît

$$D(u, z) \geqslant d \text{ dacă } |u| \leqslant c \text{ şi } |z| \leqslant b, \tag{100}$$

unde

$$D(u, z) = (\psi(u+z) - \psi(u))/z \tag{101}$$

şi c satisface

$$q = F(c) - F(-c) > 0 \tag{102}$$

(A3) $M_F(\psi^2(u)) = v < \infty$;

(A4) $M_F(\psi(u)) = 0$

(A5) $\mathbf{X}'_n\mathbf{X}_n$ nu este singulară pentu toţi $n \geqslant n_0$.

Fie $\hat{\boldsymbol{\theta}}$ orice soluţie a lui (98) şi punem:

$$\boldsymbol{\theta}^* = \mathbf{E}\boldsymbol{\theta}, \quad \hat{\boldsymbol{\theta}}^* = \mathbf{E}\hat{\boldsymbol{\theta}}, \quad \mathbf{z}_i = (\mathbf{E}')^{-1}\mathbf{x}_i \tag{103}$$

unde $\mathbf{E}$ este orice matrice $p \times p$ astfel încît $\mathbf{E}'\mathbf{E} = \mathbf{X}'\mathbf{X}$. Atunci

$$\sum_{i=1}^{n} \mathbf{z}_i\mathbf{z}'_i = \mathbf{I} \tag{104}$$

$$\sum_{i=1}^{n} \|\mathbf{z}_i\|^2 = p \tag{105}$$

şi $\hat{\boldsymbol{\theta}}^*$ este soluţia

$$\sum_{i=1}^{n} \psi(y_i - \mathbf{z}'_i\hat{\boldsymbol{\theta}}^*)\,\mathbf{z}_i = \mathbf{0}. \tag{106}$$

Are loc

TEOREMA 1 (Yohai, Maronna (1979)). *În presupunerile* (A.1)—(A 5) *dacă* p *este fixat*

i) $\|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|$ *este mărginit în probabilitate*

Scanned with OKEN Scanner

ii) *Dacă* $\lambda_1(\mathbf{X}'\mathbf{X}) \to \infty$, *atunci* $\hat{\boldsymbol{\theta}}$ *este un estimator consistent pentru* $\boldsymbol{\theta}$, *unde* $\lambda_1(\mathbf{A})$ *reprezintă cea mai mică valoare proprie a matricii* $\mathbf{A}$.

Observație. Teorema 1 arată că

$$(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})'(\mathbf{X}_n'\mathbf{X}_n)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_n = \mathbf{0}_p(1) \tag{107}$$

(unde $\mathbf{0}_p(1)$ arată mărginit în probabilitate) cînd $\hat{\boldsymbol{\theta}}$ este orice soluție a lui (98), presupunînd că ψ este monotonă, însă fără nici o presupunere asupra lui $\mathbf{X}_n$, cu excepția faptului că rang $\mathbf{X}_n = p$.

Atunci

$$\lambda_1(\mathbf{X}'\mathbf{X}) \to \infty \tag{108}$$

este o condiție suficientă pentru consistență. (108) nu poate fi slăbită deoarece este necesară pentru consistența estimatorului de C.M.M.P.

COROLARUL 1. *Fie* $\{\mathbf{x}_i\}$ *un șir de vectori aleatori p-dimensionali independenți și identic repartizați (i. i. r) astfel încît*

$$P(\mathbf{x}_i'\boldsymbol{\theta} = \mathbf{0}) < 1 \text{ pentru toți } \boldsymbol{\theta} \in R^p$$

și $\{u_i\}$ *un șir de variabile aleatoare i.i.r. cu funcția de repartiție* F, *astfel încît* $\{\mathbf{x}_i\}$ *și* $\{u_i\}$ *sînt independente. Fie*

$$y_i = \mathbf{x}_i'\boldsymbol{\theta} + u_i$$

(*ceea ce este echivalent cu presupunerea că funcția de repartiție a lui* y_i *dat* $\mathbf{x}_i$ *este* $F(y - \mathbf{x}_i'\boldsymbol{\theta})$). *Dacă sînt satisfăcute presupunerile* (A.1)—(A.5), *atunci estimatorul* $\hat{\boldsymbol{\theta}}$ *definit de* (98) *este consistent.*

TEOREMA 2. (Yohai, Marronna (1979)). *Dacă sînt satisfăcute presupunerile* (A.1)—(A.5), $p = p_n$ *depinde de* n *și*

$$\lim_{n\to\infty} p_n \max_{i \le n} \| \mathbf{z}_i \|^2 = 0 \tag{109}$$

atunci

$$p_n^{-1/2} \| \hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}^* \| = \mathbf{0}_p(1).$$

Observație. Teorema 2 tratează cazul în care p poate t inde la ∞ cu n. Folosind o condiție Lindeberg puternică, se arată că

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'(\mathbf{X}_n'\mathbf{X}_n)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})/p = \mathbf{0}_p(1).$$

Observație. Presupunem acum că $\hat{\boldsymbol{\theta}}$ este soluție a ecuației (99) cînd $\hat{\sigma} \to \sigma$ în probabilitate. Atunci au loc teoremele 1 și 2 dacă înlocuim (A1)—(A5) prin (A'1) —(A'.6), unde

(A'.1).(A.1);

(A'.2) ψ_σ satisface (A.2);

(A'.3) există $\varepsilon > 0$ astfel încît $M_F(\psi_\sigma^2((1+\varepsilon)u)) < \infty$;

Scanned with OKEN Scanner

(A'.4) există $\varepsilon > 0$ astfel încît pentru toţi λ astfel încît $|\lambda| < \varepsilon$,

$$M_F(\psi_\sigma((1+\lambda)u)) = 0;$$

(A'.5), (A.5)

(A'6) există $\varepsilon > 0$ astfel încît

$$\sup_{|\lambda|\leqslant\varepsilon, |\lambda'|\leqslant\varepsilon} M_F[(\psi_\sigma((1+\lambda)u) - \psi_\sigma((1+\lambda')u)]^2/(\lambda-\lambda')^2 \leqslant \infty.$$

Dacă $n^{1/2}(\hat{\sigma} - \sigma) = \mathbf{0}_p(1)$. (A'.4) poate fi înlocuită prin $M_F(\psi_\sigma(u)) = 0$. Argumentele în demonstraţie sînt asemănătoare cu acelea folosite de Bickel ((1975).

Au loc teoremele

TEOREMA 3 (Yohai, Maronna (1979)). *Presupunem că ψ şi F satisfac* (A.1) — (A.5) *şi următoarele condiţii*

(C.1) $$\int_{-\infty}^{\infty} [\psi(x+h) - \psi(x-h)]^2 \mathrm{d}F(x) = 0(1) \qquad \textit{cînd } h \to 0$$

şi

$$\sup_{|q|\leqslant\varepsilon, |h|\leqslant\varepsilon} \left\{ |h|^{-1} \int_{-\infty}^{\infty} [\psi(x+q+h) - \psi(x+q)]\, \mathrm{d}F(x) \right\} < \infty \textit{ pentru unii } \varepsilon > 0;$$

(C.2) *Există $A(\psi, F)$ astfel încît*

$$\int_{-\infty}^{\infty} [\psi(x+h) - \psi(x)] \mathrm{d}F(x) = hA(\psi, F) + 0(h)$$

Presupunem de asemenea

(G) $$\lim_{n\to\infty} \max_{i\leqslant n} \| \mathbf{z}_i \|^2 = 0.$$

Fie $\hat{\boldsymbol{\theta}}^*$ definit ca în (103) cu $\hat{\boldsymbol{\theta}}$ definit în (98). Atunci repartiţia lui $(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*)$ tinde către o repartiţie normală multidimensională cu vectorul valoare medie 0 şi matricea de covariante $\tau^2\mathbf{I}$ unde unde

(110) $$\tau^2 = M_F\psi^2 / A(\psi, F)^2.$$

TEOREMA 4 (Yohai, Maronna (1979)). *Presupunem că în plus faţă de* (A.1) — (A.5), *ψ şi F satisfac următoarele condiţii:*

(N.1) *ψ este de trei ori derivabilă şi* $\| \psi'''(u) \| \leqslant c$,

(N.2) $M_F \| \psi'(u) \|$ *şi* $M_F(\psi''(u))^2$ *sînt finite*

(N.3) $M_F\psi''(u) = 0$.

Presupunem de asemenea că:

(N.4) $$\lim_{n\to\infty} p_n^{3/2} \varepsilon_n = 0,$$

unde

$$\varepsilon_n = \max_{i\leqslant n} \| \mathbf{z}_i \|^2.$$

Scanned with OKEN Scanner

Fie $\mathbf{a}_n$ *orice şir din* R^{p_n}, *cu* $\|\mathbf{a}_n\| = 1$. *Atunci repartiţia statisticii* $\mathbf{a}_n'(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*)$ *tinde la* $N(0, \tau^2)$, *unde* τ^2 *este definit în* (110) (*în acest caz* $A(\psi, F) = M_F(\psi')$).

Observaţie. i) $\hat{\boldsymbol{\theta}}^*$ poatefi privit ca un estimator cu *un singur pas* cu $\hat{\boldsymbol{\theta}}^*$ ca *estimator iniţial*. După teorema 1 estimatorul iniţial are ordinul de convergenţă cerută de teorema 4.1 a lui Bickel (1975).

ii) Dacă $\hat{\boldsymbol{\theta}}$ este soluţia lui (99) cu $n^{1/2}(\hat{\sigma} - \sigma) = \mathbf{0}_p(1)$ şi (A.1) — (A.5) sînt înlocuite prin (A'.1)—(A'.6) şi condiţia (G) are loc, atunci Teorema 4.1 (Bickel (1975)) dă de asemenea condiţii suficiente pentru convergenţa statisticii $(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*)$ către repartiţia $N(\mathbf{0}, \tau^2\mathbf{I})$, unde

$$\tau^2 = M_F\psi_\sigma^2 / A(\psi_\sigma, F)^2.$$

§ 3. Estimaţii R ale coeficienţilor de regresie

3.1. Construcţia unei estimaţii a parametrului vector de regresie bazată pe alegerea convenabilă a statisticilor de rang

3.1.1. Notaţii. Presupuneri

Presupunem că pentru $N = 1, 2, \ldots$ sînt satisfăcute următoarele presupuneri

(A.1) $Y_1, Y_2, \ldots, Y_N$ sînt variabile aleatoare independente, Y_i avînd funcţia de repartiţie

$$F(y - \alpha^0 - \boldsymbol{\beta}^0\mathbf{c}^{(i)}), \qquad i = 1, 2, \ldots, N \tag{111}$$

pentru F posedînd informaţie Fisher finită, adică

$$I(F) = \int [f'(x)/f(x)]^2 f(x)\mathrm{d}x < \infty$$

cu f densitatea de repartiţie corespunzătoare lui F.

(A.2) $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ este vectorul parametrilor reali

(A.3) $\mathbf{C}_N = (c_{ji})$ este o matrice $p \times N$ cu liniile $\mathbf{c}_{(j)}$ şi coloanele $\mathbf{c}^{(i)}$, ce satisfac condiţiile

(A.3.a) $c_{ji} = c'_{ji} + c''_{ji}$, $\quad 1 \leqslant j \leqslant p$, $1 \leqslant i \leqslant N$

(A.3.b) Vectorii $\mathbf{c}'_{(j)} = (c'_{j1}, \ldots, c'_{jN})$, $j = 1, 2, \ldots, p$ satisfac

$$(\mathbf{c}'_{(j)} - \bar{c}'_j)(\mathbf{c}'_{(j)} - \bar{c}'_j)' = 0 \tag{112}$$

pentru toţi însă un număr finit de N, sau

$$(\mathbf{c}'_{(j)} - \bar{c}'_j)(\mathbf{c}'_{(j)} - \bar{c}'_j)' > 0 \tag{113}$$

pentru toţi însă un număr finit de N; mai departe

$$N^{-1}(\mathbf{c}'_{(j)} - \bar{c}'_j)(\mathbf{c}'_{(j)} - \bar{c}'_j)' \leqslant E_1 \qquad \text{pentru } N = 1, 2, \ldots \tag{114}$$

Scanned with OKEN Scanner

și dacă (113) este stisfăcută, atunci

$$\lim_{N\to\infty} \max_{1\leq i\leq N} (c'_{ji} - \bar{c}'_j)^2 \Big/ \left[\sum_{i=1}^{N} c_{ji} - \bar{c}'_j)^2\right] = 0 \tag{115}$$

unde $\bar{c}'_j = 1/N \left[\sum_{i=1}^{N} c'_{ji}\right]$ și $E_1 > 0$ este o constantă independentă de N. Presupuneri asemănătoare sînt satisfăcute de vectorii $\mathbf{c}''_j = 1, 2, \ldots, p$.

(A.3.c) Pentru toate perechile $j, l = 1, 2, \ldots, p$ au loc

$$(c'_{ji} - c'_{jk})(c'_{li} - c'_{lk}) \geqq 0$$

$$(c'_{ji} - c'_{jk})(c''_{li} - c''_{lk}) \leqq 0$$

$$(c''_{ji} - c''_{jk})(c''_{li} - c''_{lk}) \geqq 0$$

pentru toți $i, k = 1, 2, \ldots, N$; $N = 2, 3, \ldots$

(A.3.d)

$$\lim_{N\to\infty} N^{-1} (\mathbf{c}_{(l)} - \bar{c}_l)(\mathbf{c}_{(j)} - \bar{c}_j)' = \sigma_{lj}$$

$l, j = 1, 2, \ldots, p$, unde $\mathbf{\Sigma} = (\sigma_{lj})_{1\leq l,\, j\leq p} = (\boldsymbol{\sigma}^{(1)}, \ldots, \boldsymbol{\sigma}^{(k)})$ este o matrice pozitiv definită.

(A.4) $(R_1^{\beta}, R_2^{\beta}, \ldots, R_N^{\beta})$ este vectorul rangurilor ce corespunde la variabilele $Y_i - \boldsymbol{\beta}\mathbf{c}^{(i)}$, $i = 1, 2, \ldots, N$.

(A.5.) Considerăm statisticile liniare de rang

$$S_{Nj}(Y - \boldsymbol{\beta}\mathbf{c}) = N^{-1/2} \sum_{i=1}^{N} (c_{ji} - \bar{c}_j)\, a_N(R_i^{\beta}), \qquad j = 1, 2, \ldots, p \tag{116}$$

unde scorurile $a_N(i)$, $i = 1, 2, \ldots, N$ sînt generate de o funcție $\varphi(u)$, $0<u<1$ prin unul din modurile

$$a_N(i) = M\varphi(U_{Ni}) \tag{117}$$

$$a_N(i) = \varphi(i/(N+1)), \qquad i = 1, 2, \ldots, N \tag{118}$$

unde U_{Ni} este statistica de ordine de rang i într-o selecție de volum N asupra unei variabile repartizate uniform pe (0, 1).

(A.6) Funcția generatoare de scor $\varphi(u)$ nu este o funcție constantă, este nedescrescătoare și de patrat integrabilă pe (0, 1).

Observație. (114) și presupuneri analoge pentru $\mathbf{c}''_{(j)}$ implică că

$$N^{-1}(\mathbf{c}_{(j)} - \bar{c}_j)(\mathbf{c}_{(j)} - \bar{c}_j)' \leq 2\{N^{-1}(\mathbf{c}'_{(j)} - \bar{c}'_j)(\mathbf{c}'_{(j)} - \bar{c}'_j)' +$$

$$+ N^{-1}(\mathbf{c}''_{(j)} - \bar{c}''_j)(\mathbf{c}''_{(j)} - \bar{c}''_j)'\} \leq 2(E_1 + E_2) = E \text{ pentru } N = 1, 2, \ldots$$

Scanned with OKEN Scanner

Fie

$$\varphi(u, f) = -f'(F^{-1}(u))/f(F^{-1}(u)) \tag{119}$$

$$\gamma = \int_0^1 \varphi(u)\varphi(u, f)\,du \tag{120}$$

$$A^2 = \int_0^1 (\varphi(u) - \bar{\varphi})^2 du, \text{ unde } \bar{\varphi} = \int_0^1 \varphi(u)\,du \tag{121}$$

Observații. i) Modelul de regresie descris nu acoperă întreaga clasă de modele de regresie unde metoda C.M.M.P. are succes; acest fapt este datorat presupunerii (A.3.d).

ii) Nu notăm explicit dependența variabilelor în model

iii) Fără a pierde din generalitate presupunem că

$$N^{-1}(\mathbf{c}_{(j)} - \bar{c}_j)(\mathbf{c}_{(j)} - \bar{c}_j)' = 1,$$

ceea ce se obține printr-o reparametrizare proprie.

În scopul estimării parametrilor regresiei considerăm noi variabile care sînt legate de cele vechi prin

$$N^{1/2}\boldsymbol{\beta} = \Delta,\ N^{-1/2} c_{ji} = x_{ji} \quad i = 1, 2, \ldots, N,\ j = 1, 2, \ldots, p. \tag{122}$$

Astfel obținem modelul; Y_i are funcția de repartiție:

$$F(y - \alpha^0 - \Delta^0 \mathbf{x}^{(i)}), \quad i = 1, 2, \ldots, N; \tag{123}$$

proprietățile variabilelor modelului urmînd din (A.1)—(A.6). Astfel vom considera statistica

$$S_{Nj}(Y - \Delta\mathbf{x}) = \sum_{i=1}^{N} (x_{ji} - \bar{x}_j)\, a_N(R_i^{\Delta}), \quad j = 1, 2, \ldots, p, \tag{124}$$

unde $R_1^{\Delta}, \ldots, R_N^{\Delta}$ este vectorul rangurilor corespunzător la variabilele $Y_i - \Delta\mathbf{x}^{(i)}$, $i = 1, 2, \ldots, N$; considerăm posibilitatea estimării lui Δ cu ajutorul statisticii $S_{Nj}(Y - \Delta\mathbf{x})$, $j = 1, 2, \ldots, p$.

Au loc teoremele

TEOREMA 5 (Jurečková (1971a)). *Fie P_{Δ^0} repartiția probabilistică cu densitatea*

$$p_{\Delta^0} = \prod_{i=1}^{N} f(y_i - \alpha - \Delta^0 \mathbf{x}^{(i)});$$

fie

$$\|\Delta - \Delta^0\| = [(\Delta - \Delta^0)(\Delta - \Delta^0)']^{1/2}.$$

Atunci în presupunerile (A.1)— (A.6) *și sub* (122)

$$\lim_{N\to\infty} P_{\Delta^0}\{\max_{\|\Delta-\Delta^0\|<C} |S_{Nj}(Y - \Delta\mathbf{x}) - S_{Nj}(Y - \Delta^0\mathbf{x}) + \gamma(\Delta - \Delta^0)\boldsymbol{\sigma}^{(j)}| \geqq \varepsilon\} = 0 \tag{125}$$

pentru orice $\varepsilon > 0$, $c > 0$ *și* $j = 1, 2, \ldots, p$.

COROLARUL 2. *Sub presupunerile* (A.1)—(A.6) *și sub* (122)

$$\lim_{N\to\infty} P_{\Delta^0}\left\{\max_{\Delta-\Delta^0\in I}\left|\sum_{j=1}^{p} |S_{Nj}(Y-\Delta\mathbf{x})| - \sum_{j=1}^{p} |S_{Nj}(Y-\Delta^0\mathbf{x}) - \gamma(\Delta-\Delta^0)\boldsymbol{\sigma}^{(j)}|\right| \geqq \varepsilon\right\} = 0$$

pentru orice $c > 0$ *și* $\varepsilon > 0$ *și* $I = \{\mathbf{x} : \|\mathbf{x}\| \leqq c\}$.

Scanned with OKEN Scanner

3.1.2. Estimarea parametrilor $\Delta_1, \ldots, \Delta_p$.

Hájek, Šidák (1967) arată că

$$S_N(Y - \Delta^0 \mathbf{x}) = (S_{N1}(Y - \Delta^0 \mathbf{x}), \ldots, S_{Np}(Y - \Delta^0 \mathbf{x}))$$

este asimptotic repartizat

$$N(\mathbf{0}, A^2 \Sigma), \tag{127}$$

unde Σ este matricea din (A.3).

Astfel sîntem conduşi la a reprezenta parametrul necunoscut prin Δ pentru care $S_N(Y - \Delta \mathbf{x})$ este cît mai aproape posibil de zero. O posibilitate este de a reprezenta parametrul necunoscut prin mulţimea de puncte:

$$D_N = \left\{ \Delta = (\Delta_1, \Delta_2, \ldots, \Delta_p);\ \sum_{j=1}^{p} |S_{Nj}(Y - \Delta \mathbf{x})| = \min \right\}. \tag{128}$$

Mulţimea $D_n \subset E^k$ nu este vidă pentru toţi $Y_1, \ldots, Y_N$, pentru $S_{Nj}(Y - \Delta \mathbf{x})$ $(j = 1, 2, \ldots, p)$ ca funcţie de $\Delta_1, \ldots, \Delta_p$ cu $Y_1, \ldots, Y_N$ fixaţi luînd un număr finit de valori distincte.

În cele ce urmează ne ocupăm de proprietăţile asimptotice ale punctelor din D_N. Toate punctele din D_N sînt asimptotic echivalente în sensul că ele au toate aceeaşi repartiţie asimptotică.

DEFINIŢIA 1. Vom spune că fiecare punct al mulţimii D_N este asimptotic normal $(d, \mathbf{A})$ dacă există un şir de vectori aleatori $\{\tilde{\Delta}_N\}_{1 \leqslant N < \infty}$ asimptotic normal $(d, \mathbf{A})$ astfel încît

$$\lim_{N \to \infty} P\{ \sup_{\Delta \in D_N} \| \Delta - \hat{\Delta}_N \| \geqq \varepsilon \} = 0 \tag{129}$$

pentru orice $\varepsilon < 0$.

TEOREMA 6 (Jurečková (1970 a)) *Sub presupunerile* (A.1)—(A.6) *şi* (122), *fiecare punct al mulţimii* D_N *este asimptotic*

$$N(\Delta^0, \gamma^{-2} A^2 \Sigma^{-1}).$$

Observaţie. Pentru orice $Y_1, Y_2, \ldots, Y_N$ fixate considerăm următoarea funcţie de Δ:

$$h_Y(\Delta) = \sum_{j=1}^{p} |S_{Nj}(Y - \Delta^0 \mathbf{x}) - \gamma(\Delta - \Delta^0)\sigma^{(j)}|.$$

Avem $\gamma \geqslant 0$ şi dacă putem presupune că φ este astfel încît

$$\int_0^1 \varphi(u)\varphi(u, f)\, du \neq 0 \quad \text{atunci} \quad \gamma > 0$$

(acesta este cazul cînd atît $\varphi(u)$ cît şi $\varphi(u, f)$ sînt legate de repartiţii puternic unimodale). $h_Y(\Delta)$ este o funcţie convexă de Δ care are un minim unic. Punctul de minim este egal cu soluţia sistemului de ecuaţii.

$$S_{Nj}(Y - \Delta^0 \mathbf{x}) = \gamma(\Delta - \Delta^0)\, \sigma^{(j)},\ j = 1, 2, \ldots, p. \tag{130}$$

Notăm soluţia lui (130), care există şi este unică pentru orice $Y_1, \ldots, Y_N$ fixate pentru care membrii din stînga sînt bine definiţi, prin $\hat{\Delta}_N$.

Scanned with OKEN Scanner

Normalitatea asimptotică a lui $S_N(Y-\Delta^0\mathbf{x})$ implică că şirul $\{\hat{\Delta}_N\}$ este asimptotic normal

$$N(\Delta^0, \gamma^{-2} A^2 \Sigma^{-1}) \tag{131}$$

Jurečková demonstrează teorema 6 cu ajutorul lemelor

LEMA 2. *În presupunerile* (A.1)—(A.6) *şi* (122)

$$\lim_{N\to\infty} P_{\Delta^0}\left\{\sum_{j=1}^{p} |S_{Nj}(Y - \hat{\Delta}_N \mathbf{x})| \geq \varepsilon\right\} = 0 \tag{132}$$

pentru orice $\varepsilon > 0$.

COROLARUL 3. *Dacă* ε *este orice număr pozitiv, atunci*

$$\lim_{N\to\infty} P_{\Delta^0}\left\{\min_{\Delta\in E^k} \sum_{j=1}^{p} |S_{Nj}(Y - \Delta\mathbf{x})| \geq \varepsilon\right\} = 0. \tag{133}$$

LEMA 3. *Pentru orice* $c > 0$, *fie* $O_c = \{\Delta \in E^k;\ \|\Delta - \Delta^0\| \leq c\}$. *Atunci în presupunerile* (A.1) — (A.6) *şi* (122)

$$\lim_{N\to\infty} P_{\Delta^0}\{\sup_{\Delta\in D_N\cap O_c} \|\Delta - \hat{\Delta}_N\| \geq \varepsilon,\ D_N \cap O_C \neq \phi\} = 0 \tag{134}$$

pentru orice $c > 0$ *şi* $\varepsilon > 0$.

LEMA 4. *În presupunerile* (A.1) — (A.6) *şi* (122) *există* $C^* > 0$, $\delta > 0$ *şi indicele* N_0 *ce corespunde la orice* $\varepsilon > 0$ *astfel încît*

$$P_{\Delta^0}\left\{\min_{\|\Delta-\Delta^0\|\geq C^*} \sum_{j=1}^{p} |S_{Nj}(Y - \Delta\mathbf{x})| < \delta\right\} < \varepsilon \tag{135}$$

pentru orice $N > N_0$.

LEMA 5. *În presupunerile* (A.1)—(A.6) *şi* (122), *dacă*

$$\bar{O}_c = \{\Delta \in E^k;\ \|\Delta - \Delta^0\| \geq c\}$$

pentru orice $c > 0$, *atunci corespunzător la orice* $\varepsilon > 0$, *există* $C^* > 0$ *şi un întreg pozitiv* N_0, *astfel încît*

$$P_{\Delta^0}\{D_n \cap \bar{O}_{C^*} \neq \phi\} < \varepsilon \tag{136}$$

pentru orice $N > N_0$.

LEMA 6. *În presupunerile* (A.1)—(A.6) *şi* (122)

$$\lim_{N\to\infty} P_{\Delta^0}\{\sup_{\Delta\in D_N} \|\Delta - \hat{\Delta}_N\| \geq \varepsilon\} = 0 \tag{137}$$

pentru orice $\varepsilon > 0$.

Cu acestea, teorema 6, este o consecinţă a Lemei 6 şi a normalităţii asimptotice (131) a lui $\hat{\Delta}_N$.

3.1.3. Estimarea parametrului α

Presupunem că avem o regulă de alegere a unui punct unic din D_N; fie acesta $\hat{\Delta}_N$.

Adăugăm la presupunerile (A.1)—(A.6) presupunerea că $f(y)$ este simetrică faţă de zero şi $\sum_{i=1}^{N} x_{ji} = 0$, $j = 1, 2, \ldots, K$, $N = 2, 3, \ldots$.

Scanned with OKEN Scanner

Punem

(138) $\varphi^+(u,\ f) = -f'[F^{-1}((u+1)/2]/f[F^{-1}((u+1)/2)],\ 0 < u < 1$

Jurečková (1970 a) propune ca estimator pentru α

(139) $$\hat{\alpha}_N = \frac{1}{2}(\alpha_N^* + \alpha_N^{**})$$

unde

(140) $$\alpha_N^* = \sup\{\alpha;\ S_N^+(Y - \alpha - \hat{\Delta}_N \mathbf{x}) > 0\},$$
$$\alpha_N^{**} = \inf\{\alpha;\ S_N^+(Y - \alpha - \hat{\Delta}_N \mathbf{x}) < 0\}.$$

Aici $S_N^+(Y - \alpha - \hat{\Delta}\mathbf{x})$ reprezintă statistica rang

(141) $$S_N^+(Y - \alpha - \Delta\mathbf{x}) = \sum_{i=1}^{N} \operatorname{sgn}(Y_i - \alpha - \Delta\mathbf{x}^{(i)}) a_N^+(\mathscr{R}_i^{\Delta,\alpha})$$

unde $\mathscr{R}_1^{\Delta,\alpha}, \ldots, \mathscr{R}_N^{\Delta,\alpha}$ este vectorul rangurilor pentru variabilele $|Y_i - \alpha - \Delta\mathbf{x}^{(i)}|$, $i = 1, 2, \ldots N$; scorurile a_N^+ sînt generate de o funcţie de patrat integrabilă nenegativă şi nedescrescătoare $\varphi^+(u)$, $0 < u < 1$ (după (117) sau (118)).

Au loc afirmaţiile

$\bar{\Delta}_N(Y + a + \mathbf{b}\mathbf{x}) = \bar{\Delta}_N(Y) + \mathbf{b}$ pentru orice a, $\mathbf{b}$ şi repartiţia lui $\hat{\alpha}_N$ este continuă,

(142) $$P\{S_N^+(Y - a - \bar{\Delta}_N\mathbf{x}) < 0\} \leqq P\{\hat{\alpha}_N < a\} \leqq P\{S_N^+(Y - a - \hat{\Delta}_N\mathbf{x}) \leqq 0\},$$

(143) $$\lim_{N\to\infty} P_{\alpha,\Delta^\circ}\{N^{1/2}(\hat{\alpha}_N - \alpha) \leqq a\} = \lim_{N\to\infty} P_{0,0}\{S_N^+(Y - a/N^{1/2} - \bar{\Delta}_N\mathbf{x}) \leqq 0\}.$$

(144) $$\lim_{N\to\infty} P_{0,0}\{|S_N^+(Y - a/N^{1/2} - \bar{\Delta}_N\mathbf{x}) - S_N^+(Y - a/N^{1/2})| \geqq \varepsilon N^{1/2}\} = 0$$

pentru orice $\varepsilon > 0$,
α_N este asimptotic normal

(145) $$\left(\alpha,\ N^{-1}\int_0^1 [\varphi^+(u)]^2 du \cdot \left[\int_0^1 \varphi^+(u)\,\varphi^+(u, f)\,du\right]^{-2}\right).$$

3.2. Estimarea coeficienţilor de regresie prin minimizarea dispersiei rezidurilor

3.2.1. Definirea unor măsuri ale dispersiei şi proprietăţi ale estimatorilor regresiei

Jaeckel (1972) estimează coeficienţii de regresie în modelul liniar găsind acele valori ale coeficienţilor care fac rezidurile cît mai mici posibile.

Scanned with OKEN Scanner

Fie $a_N(k)$, $k = 1, 2, \ldots, N$ o mulțime de scoruri nedescrescătoare, nu toate egale ce satisfac

$$\sum_{k=1}^{N} a_N(k) = 0. \tag{146}$$

Astfel de scoruri pot fi generate de (117) sau (118).

Pentru orice $\mathbf{z} = (z_1, z_2, \ldots, z_N)$ fie $z_{(1)} \leqslant z_{(2)} \leqslant \ldots \leqslant z_{(N)}$ statisticile de ordine și definim (Jaeckel (1972))

$$D(\mathbf{z}) = \sum_{k=1}^{N} a_N(k)\, z_{(k)}.$$

După (146), $D(\mathbf{z})$ este invariantă la translație. Deoarece $D(\mathbf{z})$ are o valoare mică cînd z_i sînt apropiate unele de altele, o vom folosi ca măsură a dispersiei.

Pentru $Y_1, Y_2, \ldots, Y_N$ fixate și pentru orice β, rezidurile sînt

$$z_i = Y_i - \beta c^i \qquad i = 1, 2, \ldots, N. \tag{147}$$

Rezidurile ordonate sînt

$$z_{(k)} = Y_{i(k)} - \beta c^{i(k)}, \qquad k = 1, 2, \ldots, N$$

unde $i(k)$ este indicele observației cu residul de ordinul k (Dacă două reziduri sînt egale nu avem nici o ambiguitate în $i(k)$ însă nu în $z_{(k)}$). Dispersia reziduurilor, ca funcție de β, este

$$D(\mathbf{Y} - \beta\mathbf{C}) = \sum_{k=1}^{N} a_N(k)[Y_{i(k)} - \beta c^{i(k)}]. \tag{148}$$

Proprietățile funcției $D(\mathbf{Y} - \beta\mathbf{C})$ sînt date în

TEOREMA 7 (Jaeckel (1972)). *Pentru Y fixat, $D(\mathbf{Y} - \beta C)$ este nenegativă, continuă și convexă de β.*

Demonstrație. Fie $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_N)$ orice vector N dimensional. Fie l indicele primului $a_N(k)$ pozitiv. Atunci

$$D(\mathbf{z}) = \sum a_N(k) z_{(k)} = \sum a_N(k)[z_{(k)} - z_{(l)} \geqslant 0,$$

deoarece fiecare termen din sumă este nenegativ. Deoarece D este nenegativ și dacă z_i nu sînt toți egali, D este pozitiv.

Fie $\mathbf{p} = (p(1), p(2) \ldots, p(N))$ orice permutare a indicilor $1, 2, \ldots, N$.

Fie P mulțimea tuturor astfel de permutări. Statistica

$$D_{\mathbf{p}} = \sum a_N(k) z_{p(k)}$$

este minimizată pe P de orice permutare care aranjează z_i în ordine nedescrescătoare. Demonstrația constă în a arăta că putem trece de la orice permutare la alta care ordonează z_i printr-un șir de transpoziții, nici una scăzînd D_p.

Prin urmare putem scrie

$$D(\mathbf{Y} - \beta C) = \sum a_N(k) z_{(k)} = \max_{p \in P} \sum a_N(k) z_{p(k)},$$

Scanned with OKEN Scanner

unde acum z_i sînt rezidurile definite în (147). Pentru fiecare **p** în P avem

$$\sum a_N(k) z_{p(k)} = \sum a_N(k)[Y_{p(k)} - \beta C^{p(k)}],$$

care este o funcţie liniară de β şi prin urmare continuă şi convexă în β. Maximul unui număr finit de astfel de funcţii este continuu şi convex. Teorema este demonstrată.

Observaţie. Spaţiul lui β este împărţit într-un număr finit de submulţimi poligonale convexe, pe fiecare $D(\mathbf{Y} - \beta\mathbf{C})$ fiind o funcţie liniară de β.

Fie **E** o matrice $N \times N$ ale cărei toate intrări sînt $1/N$.

Fie $\bar{\mathbf{C}} = \mathbf{CE}$.

Cum are loc

TEOREMA 8 (Jaeckel (1972)). *Dacă* $\mathbf{C} - \bar{\mathbf{C}}$ *are rangul* K, *atunci pentru orice* D_0, *mulţimea*

$$\{\beta : D(\mathbf{Y} - \beta\mathbf{C}) \leqslant D_0\}$$

este mărginită.

$D(\mathbf{Y} - \beta\mathbf{C})$ îşi atinge minimul şi mulţimea punctelor β pentru care acesta se realizează este mărginită.

DEFINIŢIA 2 (Jaeckel(1972)). Fie β_D orice β care minimizează $D(\mathbf{Y} - \beta(\mathbf{C})$. Atunci β_D este un estimator al vectorului coeficienţilor de regresie β^0.

Cu Δ şi x_{ji} definite în (122), alegem Δ_D ca estimaţie a parametrului Δ^0 minimizînd $D(\mathbf{Y} - \Delta\mathbf{X})$. Fără a micşora din generalitate putem presupune că $\Delta^0 = \mathbf{0}$. Scriem (148) ca

$$D(\mathbf{Y} - \Delta\mathbf{X}) = \sum_{k=1}^{N} a_N(k)[Y_{i(k)} - \Delta x^{i(k)}] = \sum_{i=1}^{N} a_N(R_i^{\Delta})[Y_i - \Delta \mathbf{x}^i],$$

cu R_i^{Δ} definite ca în (124).

Derivatele parţiale ale funcţiei $D(\mathbf{Y} - \Delta\mathbf{X})$ există pentru aproape toţi Δ şi unde există sînt:

$$\frac{\partial}{\partial \Delta_j} D(\mathbf{Y} - \Delta\mathbf{X}) = \sum a_N(R_i^{\Delta})[-x_{ji})] = -\sum a_N(R_i^{\Delta})(x_{ji} - \bar{x}_j) =$$

(149)

$$= -S_{Nj}(\mathbf{Y} - \Delta\mathbf{X}), \qquad j = 1, 2, \ldots, K.$$

Astfel, acestea sînt minusul statisticilor de rang S_{Nj} definite în (124).

3.2.2. Echivalenţa asimptotică a estimaţiilor propuse de Jurečková şi Jaeckel

Jurečková alege Δ astfel încît să minimizeze $\sum |S_{Nj}(\mathbf{Y} - \Delta\mathbf{X})|$. Aceasta are drept efect apropierea tuturor $S_{Nj}(\mathbf{Y} - \Delta\mathbf{X})$ de zero.

Jaeckel alege Δ astfel încît să minimizeze $D(\mathbf{Y} - \Delta\mathbf{X})$, ceea ce face ca toate derivatele sale parţiale să fie aproape de zero. Însă derivatele

Scanned with OKEN Scanner

lui D sînt minusul statisticii S_{Nj} astfel încît aceste două procedee par a fi esenţial aceleaşi. Vom arăta că în presupunerile (A.1)—(A.6), $\{\Delta_D\}$ este echivalent cu $\{\hat{\Delta}_N\}$.

Fie

$$S_N(\mathbf{Y}) = (S_{N1}(Y),\ S_{N2}(\mathbf{Y}),\ \ldots,\ S_{Nk}(\mathbf{Y})).$$

Deoarece F este presupusă continuă, $S_{Nj}(\mathbf{Y})$ sînt bine definite cu probabilitatea unu. Definim forma pătratică

$$Q(\Delta) = \frac{1}{2}\gamma\Delta\Sigma\Delta - S_N(\mathbf{Y})\Delta + D(\mathbf{Y}).$$

Derivatele parţiale ale lui $Q(\Delta)$ sînt

$$\frac{\partial Q}{\partial \Delta_j} = \gamma\Delta\sigma^j - S_{Nj}(\mathbf{Y}). \tag{150}$$

Punctul unic în care $Q(\Delta)$ îşi atrage minimul este soluţia ecuaţiilor

$$S_{Nj}(\mathbf{Y}) = \gamma\Delta\sigma^j \qquad j = 1,\ 2,\ \ldots,\ K.$$

Notăm acest punct prin Δ_N. Acesta este acelaşi cu cel definit în (130) şi are repartiţia dată în (131).

Fie

$$B_N = \{\Delta : D(\mathbf{Y} - \Delta\mathbf{X}) \textit{ este minim}\}.$$

După teorema 8, B_N este nevidă şi mărginită. Astfel estimaţia Δ_N este orice punct al mulţimii B_N.

DEFINIŢIA 3. Cele două şiruri de vectori aleatori $\{\Delta_D\}$ şi $\{\hat{\Delta}_N\}$ se numesc asimptotic echivalente dacă distanţa dintre termenii corespunzători converg către 0 în probabilitate, adică dacă

$$\lim_{N\to\infty} P_0\{\sup_{\Delta\in B_N} \|\Delta - \Delta_N\| \geqslant r\} = 0 \tag{151}$$

pentru toţi $r > 0$.

TEOREMA 9 (Jaeckel (1972). *În presupnùerile* (A.1)—(A.6), $\{\Delta_D\}$ şi $\{\hat{\Delta}_N\}$ *sînt asimptotic echivalente. Urmează că* Δ_D *este asimptotic normal*

$$(\mathbf{0},\ \gamma^{-2}\ A^2\ \Sigma^{-1}).$$

LEMA 7. *În presupunerile teoremei 7*

$$\lim_{N\to\infty} P_0 \max_{\|\Delta\|\leqslant C} |Q(\Delta) - D(\mathbf{Y} - \Delta\mathbf{X})| \geqslant \varepsilon\} = 0$$

pentru toţi $\varepsilon > 0$ *şi* $C > 0$.

Demonstraţie. Mai întîi observăm că $Q(\mathbf{0}) = D(\mathbf{Y})$. După teorema 5 avem :

$$\lim_{N\to\infty} P_0\left\{\sup_{\|\Delta\|\leqslant C} |S_{Nj}(\mathbf{Y} - \Delta\mathbf{X}) - S_{Nj}(\mathbf{Y}) + \gamma\Delta\sigma^j| \geqslant \frac{\varepsilon}{cK^{1/2}}\right\} = 0 \tag{152}$$

pentru orice $\varepsilon > 0$, $c > 0$, și $j = 1, 2, \ldots, K$. Prin urmare după (149), (150) și (152)

$$\lim_{N\to\infty} P_0\left\{\sup_{\|\Delta\|\leqslant C}\left|\frac{\partial Q(\Delta)}{\partial \Delta_j} - \frac{\partial D(\mathbf{Y}-\Delta\mathbf{X})}{\partial \Delta_j}\right| \geqslant \frac{\varepsilon}{cK^{1/2}}\right\} = 0 \tag{153}$$

pentru $\varepsilon > 0$, $c > 0$, și $j = 1, 2, \ldots, K$.

Alegem ε și c și alegem $\Delta_0 \neq 0$ astfel ca $\|\Delta_0\| \leqslant c$. Scriem punctele de pe segmentul de la 0 la Δ_0 ca $t\Delta_0$, $0 \leqslant t \leqslant 1$. Atunci

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[Q(t\Delta_0) - D(\mathbf{Y} - t\Delta_0\mathbf{X})\right] = \sum_{j=1}^{K} \Delta_{0j}\left[\frac{\partial}{\partial \Delta_j} Q(t\Delta_0) - \frac{\partial}{\partial \Delta_j} D(\mathbf{Y} - t\Delta_0\mathbf{X})\right]. \tag{154}$$

Dacă evenimentele descrise în (153) nu au loc pentru orice j, adică dacă toate diferențele sînt mai mici decît $\varepsilon/cK^{1/2}$, atunci valoarea absolută a sumei din (154) este mai mică decît

$$\|\Delta_0\|\left(K\frac{\varepsilon^2}{c^2K}\right)^{1/2} \leqslant \varepsilon.$$

În acest caz, deoarece $Q = D$ pentru $t = 0$, și derivata diferenței lor este mai mică decît ε, avem $|Q - D| < \varepsilon$ pentru $t = 1$ sau

$$|Q(\Delta_0) - D(\mathbf{Y} - \Delta_0\mathbf{X})| < \varepsilon.$$

Dacă această inegalitate strictă are loc pentru toți Δ_0 astfel încît $|\Delta_0| \leqslant C$, atunci

$$\max_{\|\Delta\|\leqslant C} |Q(\Delta) - D(\mathbf{Y} - \Delta\mathbf{X})| < \varepsilon$$

deoarece $Q - D$ este continuă și deci trebuie să-și atingă minimul pe $\|\Delta\| \leqslant c$. Deoarece după (153) această inegalitate are loc cu probabilitate apropiată de unul, lema este demonstrată.

Cînd $K = 1$ avem:

$$D(\mathbf{Y} - \beta\mathbf{C}) = \sum a_N(k)(Y_{i(k)} - \beta C^{i(k)}),$$

unde β și fiecare c^i sînt numere. Presupunem că c^i nu sînt egale. Arătăm că dacă $a_N(k)$ sînt simetrice, β_D este „mediana ponderată" a perechilor de pante

$$(Y_j - Y_i)/(c^j - c^i).$$

Presupunem că Y_i sînt fixați. Scriem

$$D'(\mathbf{Y} - \beta\mathbf{C}) = \frac{\mathrm{d}D}{\mathrm{d}\beta} = -\sum a_N(k)c^{i(k)}.$$

Deoarece D este convexă și D' ia numai un număr finit de valori, D' este o funcție de β în scară nedescrescătoare. Deoarece, pentru β suficient de mare, avem

$$Y_j - \beta c^j < Y_i - \beta c^i$$

Scanned with OKEN Scanner

pentru toți i și j astfel încît $c^j > c^i$, urmează că

$$c^{i(1)} \geqslant c^{i(2)} \geqslant \ldots \geqslant c^{i(N)}.$$

Prin argumentul dat în teorema 7, D' își atinge maximul, care este pozitiv. Similar, pentru β suficient de mic, ordinea lui c^i este esențial inversată, astfel că D' își atinge valoarea minimă, care este negativă.

Salturile în D' apar cînd se schimbă ordinea rezidurilor. Dacă $c^j > c^i$, atunci rezidurile corespunzătoare se schimbă unul în altul cînd $\beta = \beta_{ij}$, unde

$$Y_i - \beta_{ij} c^i = Y^j - \beta_{ij} c^j, \tag{155}$$

adică cînd

$$\beta_{ij} = \frac{Y_j - Y_i}{c^j - c^i}.$$

Presupunem că toți β_{ij} sînt distincți. Acest eveniment se realizează cu probabilitatea unu deoarece F este continuă. Găsim înălțimea unuia din aceste salturi. Presupunem că pentru $\beta = \beta_{ij}$, cele două reziduri din (155) sînt statisticile de ordine de rang k și $k+1$. Atunci creșterea în D' cînd β se mișcă de la o valoare superioară la una inferioară lui β_{ij} este:

$$\begin{aligned} &- a_N(k+1)(c^i - c^j) - a_N(k)(c^j - c^i) = \\ &= [a_N(k+1) - a_N(k)](c^j - c^i). \end{aligned} \tag{156}$$

Acum definim β_D ca o anumită cuantilă a unei repartiții cu funcția de repartiție

$$G(\beta) = \frac{D'(\mathbf{Y} - \beta c) - \min D'}{\max D' - \min D'}$$

TEOREMA 10 (Jaeckel (1972)). *β_D este punctul astfel încît dacă T are funcția de repartiție*

$$P(T < \beta_D) \leqslant \frac{-\min D'}{\max D' - \min D'}$$

și

$$P(T > \beta_D) \leqslant \frac{\max D'}{\max D' - \min D'}.$$

Dacă $a_N(k)$ sînt simetrice, adică dacă

$$a_N(k) = -a_N(N+1-k), \qquad k = 1, 2, \ldots, N,$$

atunci β_D este mediana lui G.

Demonstrație. D este minimizată în β_D dacă $D' \leqslant 0$ pentru $\beta < \beta_D$ și $D' \geqslant 0$ pentru $\beta > \beta_D$. Atunci prima aserțiune urmează din defi-

niţia lui G. Dacă $a_N(k)$ sînt simetrice, $\max D' = \min D'$, de unde rezultă a doua aserţiune.

COROLARUL 4. *Dacă $a_N(k)$ sînt scorurile Wilcoxon*

$$a_N(k) = \frac{k}{N+1} - \frac{1}{2}, \qquad k = 1, 2, \ldots, N$$

atunci β_D este mediana mulţimii pantelor

$$\beta_{ij} = \frac{Y_j - Y_i}{c^j - c^i}$$

pentru $c^j > c^i$. În acest caz dispersia asimptotică a lui $N^{1/2}\beta_D$ (dacă au loc presupunerile teoremei 9) este

$$1/12\,\sigma^2 \left\{\int f^2(x)\,dx\right\}^2,$$

unde

$$\sigma^2 = \lim_{N\to\infty} \frac{1}{N} \sum (c^j - \bar{c})^2$$

şi f este densitatea lui F.

Demonstraţie. Dacă $a_N(k)$ sînt definite ca mai înainte, toate diferenţele între scorurile consecutive sînt egale, astfel că înălţimea salturilor în (156) depind numai de $c^j - c^i$. Prin urmare $G(\beta)$ are salturi în β_{ij} cu înălţimea proporţională cu $c^j - c^i$. După teorema 10, jumătate din ponderi (înălţimile salturilor lui G) cad de o parte şi alta a lui β_D.

Dispersia asimptotică este o consecinţă a teoremei 9 cu $\varphi(u) = u - \frac{1}{2}$.

3.3. Estimarea coeficienţilor de regresie prin minimizarea funcţiei de dispersie definită de diferenţa ponderată a lui Gini

3.3.1. Definiţia şi proprietăţile funcţiei de dispersie

Considerăm modelul liniar

(157) $$\mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

unde

$\mathbf{Y} = (Y_1, \ldots, Y_n)'$ este un vector $n \times 1$,

$\mathbf{1}$ este vectorul $n \times 1$ cu fiecare element egal cu unu,

$\mathbf{X} = (x_{ij})$ este o matrice $n \times p$

$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ este vectorul parametrilor şi

$\mathbf{e} = (e_1, \ldots, e_n)'$ este vectorul erorilor.

Scanned with OKEN Scanner

Presupunem că $e_1, e_2, \ldots, e_n$ sînt variabile aleatoare i.i.r avînd densitatea de repartiţie f. Rezidurile sînt date de $\mathbf{Z} = (Z_1, \ldots, Z_n)'$ unde

$$\mathbf{Z} = \mathbf{Z}(\boldsymbol{\beta}) = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}.$$

Sievers (1979) estimează parametrii modelului (157) prin minimizarea funcţiei de dispersie definită de diferenţa medie ponderată a lui Gini.

DEFINIŢIA 4. Funcţia de dispersie

$$D = D(\boldsymbol{\beta}) = \sum_{i<j} b_{ij} |Z_i - Z_j|, \tag{158}$$

unde $b_{ij} \geqslant 0$, $1 \leqslant i < j \leqslant n$ este o mulţime de ponderi date., poartă numele de diferenţa medie ponderată a lui Gini.

Fiecare pereche de reziduri este comparată prin diferenţa absolută şi ponderea căreia i se ataşează reflectă importanţa comparării. Notăm că ponderile pot depinde de $\mathbf{X}$. Este posibil să avem unele ponderi egale cu zero şi aceasta înlătură din consideraţie unele perechi de comparaţii. Cazul particular al ponderilor egale, $b_{ij} = 1$, dă diferenţa medie a lui Gini.

PROPOZIŢIA 4. (Hettmansperger, McKean (1978)). *Funcţia de dispersie* (158) *este echivalentă cu dispersia* (148) *cu scorurile Wilcoxon.*

Funcţia de dispersie D poate fi exprimată şi în altă formă.

Extindem definiţia ponderilor b_{ij} la toţi indicii $i, j = 1, \ldots, n$ folosind $b_{ji} = b_{ij}$ şi $b_{ii} = 0$. Atunci, folosind

$$|v| = v \operatorname{sgn}(v)$$

găsim

$$D = \sum_{i=1}^{n} B_i Z_i \tag{158'}$$

cu

$$B_i = \sum_{j \neq i} b_{ij} \operatorname{sgn}(Z_i - Z_j), \qquad i = 1, 2, \ldots, n.$$

Coeficienţii B_i sînt aleatori cu B_i depinzînd de rangul lui Z_i. În cazul particular în care $b_{ij} \equiv 1$, $B_i = 2R_i - (n+1)$.

O altă funcţie de dispersie care este similară cu D este

$$D^* = \sum_{i<j} b_{ij} |Z_i - Z_j|, \tag{158''}$$

în care ponderile corespund la rezidurile ordonate $Z_{(1)} \leqslant \ldots \leqslant Z_{(n)}$
Avem :

$$D^* = \sum_{i=1}^{n} B^*_{R_i} Z_i = \sum_{i=1}^{n} B_i^* Z_{(i)},$$

unde

$$B_i^* = \sum_{j=1}^{i-1} b_{ij} - \sum_{j=i+1}^{n} b_{ij}.$$

Scanned with OKEN Scanner

PROPOZIȚIA 5 *Funcția de dispersie D^* este echivalentă cu funcția de dispersie* (148) *cu B_i^* ca funcție scor.*

Dacă $b_{ij} \equiv 1$, atunci $D = D^*$. În general ele nu sînt egale. Aceasta arată că ponderile folosite în D servesc unui scop diferite de funcția scor folosită în 3.2.1.

DEFINIȚIA 5 (Sievers (1979)). Vom numi estimație a parametrului $\boldsymbol{\beta}$ punctul din spațiul parametrilor care minimizează funcția de dispersie $D(\boldsymbol{\beta})$ din (158).

PROPOZIȚIA 6. *Funcția de dispersie* (158) *este nenegativă, liniară și convexă.*

Pentru determinarea estimației pot fi folosite diferite metode, incluzînd algoritmi de programare liniară. În general soluția nu este unică. Totuși în anumite condiții, diametrul mulțimii soluțiilor tinde asimptotic la zero.

Derivatele parțiale ale lui D vor fi (aproximativ) egale cu zero în minim. Folosind forma (158'), aceste derivate sînt

$$\partial D/\partial \beta_k = -\sum_{i=1}^{n} B_i x_{ik} \tag{159}$$

pentru $k = 1, \ldots, p$ în punctul $\boldsymbol{\beta}$ unde ele există. O altă formă a derivatelor, scriind

$$D = \sum_{i<j} b_{ij} \operatorname{sgn}(Z_j - Z_i)(Z_j - Z_i)$$

este

$$\begin{aligned} \partial D/\partial \beta_k &= -\sum_{i<j} b_{ij} \operatorname{sgn}(Z_j - Z_i)(x_{jk} - x_{ik}) = \\ &= -2\sum_{i<j} b_{ij}(x_{jk} - x_{ik})\,\Phi(Z_i, Z_j) + \sum_{i<j} b_{ij}(x_{jk} - x_{ik}), \end{aligned} \tag{159'}$$

unde $\Phi(u, v) = (\operatorname{sgn}(v - u) + 1)/2 = 0,\ 1/2, 1$ după cum $u > v$, $u = v$, $u < v$.

Observații. i) Din această formă urmează că derivatele depind de rangurile ordonate ale rezidurilor. Ele conțin un tip general de variabilă aleatoare de forma

$$\sum_{i<j} a_{ij}\,\Phi(Z_i, Z_j).$$

ii) (159') se mai poate scrie

$$\partial D/\partial \beta_k = \sum_{i<j} b_{ij} |x_{jk} - x_{ik}| \operatorname{sgn}(x_{jk} - x_{jk}) \operatorname{sgn}(Z_j - Z_i) \tag{159''}$$

pentru $k = 1, \ldots, p$. Aceasta este variabila tau ponderată a lui Kendall. Astfel cînd derivatele parțiale sînt zero, rezidurile sînt necorelate cu variabilele independente în acest sens neparametric.

Scanned with OKEN Scanner

3.3.2. O clasă generală de variabile independente

Pentru fiecare $k = 1, \ldots, p$, fie dată mulţimea de constante $\{a_{ij}(k) : 1 \leqslant i < j \leqslant n\}$. Fie

$$a_{i.}(k) = \sum_{j=i+1}^{n} a_{ij}(k) \qquad \text{pentru } i = 1, \ldots, n-1$$

$$a_{.j}(k) = \sum_{i=1}^{j-1} a_{ij}(k) \qquad \text{pentru } j = 2, \ldots, n$$

$$a_{n.}(k) = 0, \; a_{.1}(k) = 0, \; a_{..}(k) = \sum_{i<j} a_{ij}(k)$$

$$A_i(k) = a_{.i}(k) - (a_{i.}k)$$

Pentru studiu asimptotic este necesar un astfel de şir de constante, indexate de $n = 1, 2, \ldots$. În cele ce urmează, această dependenţă de n va fi suspendată din notaţie.

Presupunerea (A1). Pentru fiecare $k = 1, \ldots, p$

$$\frac{\sum_{i=1}^{n} A_i^2(k)}{\max_{1 \leqslant i \leqslant n} A_i^2(k)} \to \infty \qquad \text{cînd } n \to \infty$$

Presupunerea (A2). Pentru fiecare $k = 1, \ldots, p$

$$\frac{\sum_{i<j} a_{ij}^2(k)}{\sum_{i=1}^{n} A_i^2(k)} \to 0 \quad \text{cînd } n \to \infty.$$

Definim variabilele alcatoare

$$T_k = T_k(\boldsymbol{\beta}) = \sum_{i<j} a_{ij}(k)\, \Phi(Z_i, Z_j), \tag{160}$$

pentru $k = 1, \ldots, p$, unde $\Phi(u, v) = 0, 1/2, 1$ după cum $u > v$, $u = v$, $u < v$.

Fie

$$\mathbf{T} = \mathbf{T}(\boldsymbol{\beta}) = (T_1, \ldots, T_p)'$$

vectorul $p \times 1$ de astfel de variabile. Notăm că acest tip de variabilă aleatoare apare în derivatele funcţiei în (159′) cu

$$a_{ij}(k) = b_{ij}(x_{ik} - x_{ik})$$

corespunzători. Avem :

$$\partial D / \partial \beta_k = -2T_k(\boldsymbol{\beta}) + a_{..}(k). \tag{161}$$

Scanned with OKEN Scanner

Pentru a considera repartiţia asimptotică a statisticii $\mathbf{T}(\boldsymbol{\beta})$, folosim notaţiile :

Specificăm un şir de valori ale parametrului, contigui cu 0, presupunînd $\boldsymbol{\beta} = \boldsymbol{\Delta}/\sqrt{n}$, unde $\boldsymbol{\Delta}' = (\Delta_1, \ldots, \Delta_p)$ este un vector fixat.

Considerăm matricea centrată :

$$\mathbf{X}_c = (\mathbf{I}_n - (1/n)\mathbf{J}_n)\,\mathbf{X},$$

unde $\mathbf{I}_n$ este matricea identitate $n \times n$ şi

$\mathbf{J}_n$ este matricea de unu $n \times n$.

Fie $\bar{x}_k$ media coloanei k a matricei $\mathbf{X}$,

$\mathbf{A}_n$ o matrice $n \times p$ al cărui element (i, k) este egal cu $A_i(k)$,

$$i = 1, \ldots, n, \quad k = 1, \ldots, p, \text{ şi } \mathbf{V}_n = \mathbf{A}_n'\mathbf{A}_n,$$

$\mathbf{a}_{\cdot\cdot}' = (a_{\cdot\cdot}(1), \ldots, a_{\cdot\cdot}(p))$ şi

$$\boldsymbol{\mu}_n' = (\mu_n(1), \ldots, \mu_n(p)) = \left(\int f^2\right)\mathbf{A}_n'\,\mathbf{X}_c\boldsymbol{\beta} + (1/2)\,\mathbf{a}_{\cdot\cdot}.$$

Presupunerea (A3). Pentru $k = 1, \ldots, p$

$$\max_{1 \le i \le n} \frac{|x_{ik} - \bar{x}_k|}{\sqrt{n}} \to 0 \text{ cînd } n \to \infty$$

Presupunerea (A4)

$$(1/n)\mathbf{X}_c'\mathbf{X}_c \to \boldsymbol{\Sigma} \quad \text{cînd } n \to \infty$$

unde $\boldsymbol{\Sigma}$ este o matrice pozitiv definită.

Presupunerea (A5). Există un şir de constante $\{\gamma_n\}$ astfel încît

$$\gamma_n\mathbf{V}_n \to \mathbf{V} \quad \text{cînd } n \to \infty$$

unde $\mathbf{V}$ este o matrice pozitiv definită $p \times p$.

TEOREMA 11 (Sievers (1979)). *Dacă sînt satisfăcute condiţiile — densitatea erorilor f este absolut continuă şi*

$$\int (f'/f)^2 \,\mathrm{f}\,dx < \infty$$

— (A1) —(A5)

şi dacă $\boldsymbol{\beta} = \boldsymbol{\Delta}/\sqrt{n}$, atunci

$$\gamma_n^{1/2}(\mathbf{T}(\mathbf{0}) - \boldsymbol{\mu}_n) \xrightarrow[n \to \infty]{\mathscr{R}} N(\mathbf{0}, (1/12)\mathbf{V})$$

Demonstraţie. Fie $W_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$, $i = 1, \ldots, n$ şi $\bar{W} = \sum W_i/n$. Fie $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ un vector fixat şi fără a pierde din generalitate în cele ce urmează luăm $\boldsymbol{\theta}'\boldsymbol{\theta} = 1$. Considerăm combinaţia liniară

$$U = U(\boldsymbol{\beta}) = \boldsymbol{\theta}'\mathbf{T}(\boldsymbol{\beta}) = \sum_{i<j} h_{ij}\,\Phi(Z_i, Z_j),$$

Scanned with OKEN Scanner

unde

$$h_{ij} = \theta_1 a_{ij}(1) + \ldots + \theta_p a_{ij}(p), \quad 1 \leqslant i < j \leqslant n.$$

Fie

$$h_{i.} = \sum_{j=i+1}^{n} h_{ij}, \; h_{.j} = \sum_{i=1}^{j-1} h_{ij}, \; h_{n.} = 0, \; h_{.1} = 0, \; h_{..} = \sum_{i<j} h_{ij},$$

$$H_i = h_{.i} - h_{i.} = \theta_1 A_i(1) + \ldots + \theta_p A_i(p).$$

Teorema este demonstrată dacă arătăm că $U(\boldsymbol{\theta})$ este asimptotic normal cu media

$$\sum_{i=1}^{n} H_i(w_i - \bar{w}) \left(\int f^2\right) + (h../2)$$

și dispersia

$$\sum_{i=1}^{n} H_i^2/12$$

pentru orice alegere a lui $\boldsymbol{\theta}$.

Prima presupunere cere $\max\limits_{1 \leqslant i \leqslant n} (w_i - \bar{w})^2 \to 0$. Aceasta urmează din (A3). A doua presupunere cere ca $\sum\limits_{i} (w_i - \bar{w})^2 \to \sigma_w^2 > 0$. Aceasta urmează din (A4) folosind $\sum\limits_{i}(w_i - \bar{w})^2 = \boldsymbol{\theta}' \mathbf{X}_c' \mathbf{X}_c \boldsymbol{\theta}/n$.

A treia presupunere cere $\sum\limits_{i} H_i^2 / \max\limits_{i} H_i \to \infty$.

Putem scrie

$$\frac{\sum\limits_{i} H_i^2}{\max\limits_{i} H_i^2} = \frac{\sum\limits_{i} [\theta_1 A_i(1) + \ldots + \theta_p A_i(p)]^2}{\sum\limits_{i} A_i^2(1) + \ldots + \sum A_i^2(p)} \times$$

$$\times \frac{\sum\limits_{i} A_i^2(1) + \ldots + \sum\limits_{i} A_i^2(p)}{\max\limits_{i} [\theta_1 A_i(1) + \ldots + \theta_p A_i(p)]^2} = F_1 \cdot F_2 \; (\textit{notație}).$$

Avem

$$F_1 = \frac{\boldsymbol{\theta}' \mathbf{V}_n \boldsymbol{\theta}}{\mathrm{Ur}(\mathbf{V}_n)} \geqslant \max_{\boldsymbol{\theta} \ni \boldsymbol{\theta}'\boldsymbol{\theta} = 1} \left[\frac{\boldsymbol{\theta}' \mathbf{V}_n \boldsymbol{\theta}}{\mathrm{Ur}\, \mathbf{V}_n} \right] = \frac{\lambda_{1n}}{\lambda_{1n} + \ldots + \lambda_{pn}},$$

unde $\lambda_{1n} \leqslant \lambda_{2n} \leqslant \ldots \leqslant \lambda_{pn}$ sînt valorile proprii ordonate ale lui $\mathbf{V}_n$. Fie $\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_p$ valorile proprii ordonate ale lui $\mathbf{V}$. Atunci după (A5), cînd $n \to \infty$,

$$\frac{\lambda_{1n}}{\lambda_{1n} + \ldots + \lambda_{pn}} = \frac{\gamma_n \lambda_{1n}}{\gamma_n \lambda_{1n} + \ldots + \gamma_n \lambda_{pn}} \to \frac{\lambda_1}{\lambda_1 + \ldots + \lambda_n} > 0$$

și aceasta arată că F_1 este mărginit de zero.

Din inegalitatea Cauchy-Schwarz

$$F_2 \geqslant \frac{\sum_i A_i^2(1) + \ldots + \sum_i A_i^2(p)}{\theta'\theta \max_i [A_i^2(1) + \ldots + A_i^2(p)]} \geqslant \frac{\sum_i A_i^2(1) + \ldots + \sum A_i^2(p)}{p \max \{A_i^2(1), \ldots, A_i^2(p)\}} \geqslant$$

$$\geqslant \frac{1}{p} \min \left\{ \frac{\sum_i A_i^2(1)}{\max_i A_i^2(1)}, \ldots, \frac{\sum_i A_i^2(p)}{\max_i A_i^2(p)} \right\}.$$

Atunci, după (A1) urmează că $F_2 \xrightarrow[n\to\infty]{} \infty$. Comportarea asimptotică a statisticilor F_1 și F_2 arată că a treia presupunere are loc.

A patra presupunere cere ca

$$\sum_{i<j} h_{ij}^2 / \sum_i H_i^2 \to 0.$$

Acum

$$\frac{\sum_{i<j} h_{ij}^2}{\sum_i H_i^2} = \frac{\sum_{i<j} [\theta_1 a_{ij}(1) + \ldots + \theta_p a_{ij}(p)]^2}{\sum_i A_i^2(1) + \ldots + \sum_i A_i^2(p)} \cdot \frac{\sum_i A_i^2(1) + \ldots + \sum_i A_i^2(p)}{\sum_i [\theta_1 A_i(1) + \ldots + \theta_p A_i(p)]^2} =$$

$$= G_1 G_2 \ (notație).$$

Din inegalitatea Cauchy-Schwarz,

$$G_1 \leqslant \frac{\theta'\theta [\sum_{i<j} a_{ij}^2(1) + \ldots + \sum_{i<j} a_{ij}^2(p)]}{\sum_i A_i^2(1) + \ldots + \sum_i A_i^2(p)} \leqslant$$

$$\leqslant \frac{\sum_{i<j} a_{ij}^2(1)}{\sum_i A_i^2(1)} + \ldots + \frac{\sum_{i<j} a_{ij}^2(p)}{\sum_i A_i^2(p)}.$$

Atunci cu presupunerea A_1, urmează că $G_1 \to 0$ cînd $n \to \infty$. Acum $G_2 = 1/F_1$ și am arătat că F_1 este mărginită de zero. Astfel G_2 este mărginită și din comportarea lui G_1, a patra presupunere are loc.

3.3.3. Liniaritatea asimptotică.

Arătăm că $\mathbf{T}(\beta)$ este asimptotic liniară.
Fie $\Delta = (\Delta_1, \ldots, \Delta_p)'$ un vector $p \times 1$ și $c > 0$ dat. Fie

$$\mathcal{D} = \{\Delta : -c \leqslant \Delta_k \leqslant c, \quad k = 1, \ldots, p\}.$$

Fie $\mathbf{C}_n$ o matrice $p \times p$ avînd elementul (k, l) dat de

$$c_{kl} = -\left(\int f^2\right) \sum_{i<j} a_{ij}(k)\,(x_{jl} - x_{il}).$$

Fie

$$\mathbf{R}(\Delta) = n^{-3/2}[\mathbf{T}(\Delta/\sqrt{n}) - \mathbf{T}(0) - \mathbf{C}\,(\Delta/\sqrt{n})]. \tag{162}$$

Scanned with OKEN Scanner

Fie

$$G(y) = P(e_1 - e_2 \leqslant y)$$

funcţia de repartiţie a diferenţei variabilelor independente, fiecare cu densitatea f.

Fie $\| \cdot \|$ distanţa Euclidiană.

Presupunerea (A6). Funcţia de repartiţie G are densitatea $g = G'$ şi $g(y)$ este continuă în $y = 0$.

Presupunerea (A7). Pentru fiecare $k = 1, \ldots, p$

$$\frac{\sum_{i<j} a_{ij}^2(k)}{C_n^2} \text{ este mărginită cînd } n \to \infty.$$

LEMA 8 (Sievers (1979)). *Dacă sînt satisfăcute presupunerile* (A2), (A4), (A6) şi (A7) *şi dacă* $\boldsymbol{\beta} = \mathbf{0}$, atunci

$$\mathbf{R}(\Delta) \xrightarrow{P} \mathbf{0}$$

uniform în $\Delta \in \mathscr{D}$ (*adică, pentru toţi* $\varepsilon > 0$ şi $\delta > 0$, *există* N *astfel încît* $P(\| \mathbf{R}(\Delta) \| \geqslant \varepsilon) \leqslant \delta$ *pentru toţi* $n \geqslant N$ *şi toţi* $\Delta \in \mathscr{D}$).

Demonstraţie. Este suficient să arătăm că rezultatul are loc pentru fiecare coordonată a lui $\mathbf{R}(\Delta)$. Vom considera prima coordonată şi o vom nota cu $\mathbf{R}_1(\Delta)$. Cu

$$t_{ij}(\Delta) = \Delta_1(x_{j1} - x_{i1})/\sqrt{n} + \ldots + \Delta_p(x_{jp} - x_{ip})/\sqrt{n},$$

$\mathbf{R}_1(\Delta)$ poate fi scrisă ca

$$R_1(\Delta) = n^{-3/2}\left[\sum_{i<j} a_{ij}(1)\, w_{ij} + \left(\int f^2\right) \sum_{i<j} a_{ij}(1) t_{ij}(\Delta) \right],$$

unde:

$$w_{ij} = \phi(Z_i, Z_j) - \phi(Y_i, Y_j) =$$

$$= \begin{cases} 1, & t_{ij}(\Delta) < Y_j - Y_i < 0 \\ -1, & 0 < Y_j - Y_i < t_{ij}(\Delta) \\ 0, & \text{în rest.} \end{cases}$$

Putem avea W_{ij} egal cu $\pm 1/2$ în cazul în care avem date egale; vom ignora acest caz deoarece aceste evenimente au probabilitatea de apariţie zero.

În amîndouă cazurile $t_{ij}(\Delta) > 0$ şi $t_{ij}(\Delta) < 0$,

$$M(w_{ij}) = G(0) - G(t_{ij}(\Delta)) = -t_{ij}(\Delta) g(\xi_{ij}(\Delta)),$$

unde

$$|\xi_{ij}(\Delta)| \leqslant |t_{ij}(\Delta)|.$$

Scanned with OKEN Scanner

Atunci folosind $g(0) = \int f^2$,

$$M(R_1(\Delta)) = n^{-3/2} \sum_{i<j} a_{ij}(1)\, t_{ij}(\Delta)[g(0) - g(\xi_{ij}(\Delta))].$$

Din presupunerile (A3) și (A6) urmează că pentru orice $\varepsilon > 0$,

$$\max_{i<j} |\, g(0) - g(\xi_{ij}(\Delta))\,| < \varepsilon$$

uniform în $\Delta \in \mathscr{D}$ pentru n suficient de mare. Notînd că

$$\sum_{i<j} (x_{jk} - x_{ik})(x_{jl} - x_{il})/C_n^2 = \sum_{i} (x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l)/n,$$

scriem

$$\sum_{i<j} t_{ij}^2(\Delta)/C_n^2 = \Delta' \mathbf{X}_c' \mathbf{X}_c \Delta / n.$$

Atunci, folosind inegalitatea Cauchy

$$|\, M(R_1(\Delta))\,| \leqslant \varepsilon n^{-2} C_n^2 \{\sum_{i<j} a_{ij}^2(1)/C_n^2\}^{1/2} \leqslant \{\Delta' \mathbf{X}_c' \mathbf{X} \Delta / n\}^{1/2}.$$

Cu presupunerile (A4) și (A7) și din faptul că ε este arbitrar, trebuie ca $M(R_1(\Delta)) \to 0$ cînd $n \to \infty$, uniform în $\Delta \in \mathscr{D}$.

Dispersia lui $R_1(\Delta)$ este

$$D^2(R_1(\Delta)) = n^{-3} \sum_{i<j} a_{ij}^2(1) D^2(W_{ij}) +$$

$$+ n^{-3} \sum_{\substack{i<j \\ (i,j)\neq(k,l)}} \sum_{k<l} a_{ij}(1)\, a_{kl}(1) \operatorname{cov}(W_{ij}, W_{kl}).$$

Folosind presupunerile (A3) și (A7), urmează că $D^2(R_1(\Delta)) \to 0$ cînd $n \to \infty$ uniform în $\Delta \in \mathscr{D}$. Cu media și dispersia tinzînd către zero, lema este demonstrată.

Observație. Lema arată că $\mathbf{T}(\boldsymbol{\beta})$ poate fi aproximat asimptotic printr-o funcție liniară $\mathbf{T}(\mathbf{0}) + C_n\boldsymbol{\beta}$ pentru $\boldsymbol{\beta}$ suficient de apropiat de zero. Totuși rezultatul nu este puternic pentru aplicații. Un rezultat mai puternic este dat în următoarea teoremă.

TEOREMA 12. (Sievers (1979)). *Dacă sînt satisfăcute presupunerile* (A3), (A4), (A6) *și* (A7) *și dacă* $\boldsymbol{\beta} = \mathbf{0}$, *atunci*

$$\sup_{\Delta \in \mathscr{D}} \| \mathbf{R}(\Delta) \| \xrightarrow[n\to\infty]{P} \mathbf{0}.$$

Demonstrație. Este suficient să arătăm că rezultatul are loc pentru fiecare coordonată a lui $\mathbf{R}(\Delta) \cdot R_1(\Delta)$ și $t_{ij}(\Delta)$ sînt date în lema anterioară.

Fie $\varepsilon > 0$, $\varepsilon' > 0$ dați.

Scanned with OKEN Scanner

Din presupunerile (A4) și (A7) există marginea B_0 astfel încît

$$(1/2)\ [\sum_{i<j} a_{ij}^2(1)/C_n^2]^{1/2} \left\{ \sum_{k=1}^{p} \left[\sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2/n \right]^{1/2} \right\} \leqslant B_0.$$

Alegem δ astfel încît $3g(0)\ B_0\delta < \varepsilon/3$. Atunci partiționăm $\mathscr{D}$ în submulțimile închise $\mathscr{D}_1, \ldots, \mathscr{D}_M$ astfel încît $\Delta, \Delta' \in \mathscr{D}_m$ implică $|\Delta_k - \Delta'_k| \leqslant \delta$ pentru toți $k = 1, \ldots, p$ și $m = 1, \ldots, M$ și că $\mathscr{D} = \bigcup_{m=1}^{M} \mathscr{D}_m$. Fie

$$t_{ij}^U(m) = \max_{\Delta \in \mathscr{D}_m} t_{ij}(\Delta) \text{ și } t_{ij}^L(m) = \min_{\Delta \in \mathscr{D}_m} t_{ij}(\Delta).$$

Atunci din presupunerile (A4) și (A7),

$$n^{-3/2} \sum_{i<j} a_{ij}(1)\, |t_{ij}^U(m) - t_{ij}^L(m)| \leqslant B_0\delta.$$

Pentru $1 \leqslant i < j \leqslant n$ și $m = 1, \ldots, M$ definim variabilele aleatoare

$$S_{ij}(m) = \begin{cases} 1, & t_{ij}^L(m) - Y_i - Y_j \leqslant t_{ij}^U(m) \\ 0, & \text{în rest} \end{cases}$$

și

$$Q_m = n^{-3/2} \sum_{i<j} a_{ij}(1) S_{ij}(m).$$

Dacă sînt adevărate presupunerile (A3), (A4), (A6), (A7) și $\boldsymbol{\beta} = \mathbf{0}$, obținem

$$M(Q_m) = n^{-3/2} \sum_{i<j} a_{ij}(1)\, [G(t_{ij}^U(m)) - G(t_{ij}^L(m))] \leqslant$$

$$\leqslant 2g(0) B_0 \delta$$

pentru toți $m = 1, \ldots, M$ și pentru n suficient de mare.

De asemenea

$$D^2(Q_m) \to 0 \text{ cînd } n \to \infty$$

pentru toți $m = 1, \ldots, M$.

Pentru fiecare $m = 1, \ldots, M$ alegem nu punct $\Delta_m \in \mathscr{D}_m$. Atunci

$$\sup_{\Delta \in \mathscr{D}_m} n^{-3/2} |T_1(\Delta/\sqrt{n}) - T_1(\Delta_m/\sqrt{n})| \leqslant Q_m$$

pentru fiecare $m = 1, \ldots, M$. După lema 8, pentru fiecare $m = 1, \ldots, M$ și pentru n suficient de mare

$$P(|R_1(\Delta_m)| \geqslant \varepsilon/3) \leqslant \varepsilon'/2M.$$

Scanned with OKEN Scanner

Ţinînd seama de rezultatele obţinute, pentru fiecare $m = 1, \ldots, M$ şi pentru n suficient de mare

$$\sup_{\Delta \in \mathcal{D}_m} |R_1(\Delta) - R_1(\Delta_m)| \leqslant \sup_{\Delta \in \mathcal{D}_m} n^{-3/2} |T_1(\Delta/\sqrt{n}) - T_1(\Delta_m/\sqrt{n})| +$$

$$+ \sup_{\Delta \in \mathcal{D}_m} n^{-3/2} \left(g(0) \sum_{i<j} a_{ij}(1) |t_{ij}(\Delta) - t_{ij}(\Delta_m)|\right) \leqslant$$

$$\leqslant Q_m + g(0) B_0 \delta =$$

$$= Q_m - M(Q_m) + M(Q_m) + g(0) B_0 \delta \leqslant$$

$$\leqslant Q_m - M(Q_m) + \varepsilon/3.$$

Pentru fiecare $m = 1, \ldots, M$ şi pentru fiecare n suficient de mare

$$P(\sup_{\Delta \in \mathcal{D}_m} |R_1(\Delta)| \geqslant \varepsilon) \leqslant$$

$$\leqslant P(\sup_{\Delta \in \mathcal{D}_m} |R_1(\Delta) - R_1(\Delta_m)| + R_1(\Delta_m) \leqslant \varepsilon) \leqslant$$

$$\leqslant P(Q_m - M(Q_m) + \varepsilon/3 + R_1(\Delta_m) \geqslant \varepsilon) \leqslant$$

$$\leqslant P(Q_m - M(Q_m) \geqslant \varepsilon/3) + P(|R_1(\Delta_m)| \geqslant \varepsilon/3) \leqslant$$

$$\leqslant (9/\varepsilon^2) D^2(Q_m) + \varepsilon'/2M \leqslant$$

$$\leqslant \varepsilon'/2M + \varepsilon'/2M = \varepsilon'/M.$$

În final, pentru n suficient de mare

$$P(\sup_{\Delta \in \mathcal{D}_m} |R_1(\Delta)| \geqslant \varepsilon) \leqslant \sum_{m=1}^{M} P(\sup_{\Delta \in \mathcal{D}_m} |R_1(\Delta)| \geqslant \varepsilon) \leqslant \sum_{m=1}^{M} \varepsilon'/M = \varepsilon';$$

teorema este demonstrată.

3.3.4. Repartiţia lui $\hat{\beta}$

Estimaţia parametrului β a fost definită ca fiind punctul $\hat{\beta}_n$ din spaţiul parametrilor care minimizează funcţia de dispersie $D(\beta)$ din (158). Mulţimea soluţiilor acestei probleme de minimizare este mărginită după cum rezultă din

LEMA 9. *Dacă matricea centrată* $\mathbf{X}_c$ *este de rang maxim* p, *atunci* $\{\beta : D(\beta) \leqslant D_0\}$ *este mărginită pentru orice număr* D_0.

Demonstraţia este aceeaşi cu a teoremei 8.

Fie $\hat{\Delta}_n = \sqrt{n}\hat{\beta}_n$ şi

$$D^*(\Delta) = (1/n) D(\Delta/\sqrt{n})$$

Atunci $\hat{\Delta}_n$ minimizează $D^*(\Delta)$.

Scanned with OKEN Scanner

Folosim corespondenţa

$$a_{ij}(k) = b_{ij}(x_{jk} - x_{ik}). \tag{163}$$

Atunci din (161) vectorul derivatelor parţiale este

$$\frac{\partial D^*(\Delta)}{\partial \Delta} = n^{3/2}[-2\mathbf{T}(\Delta/\sqrt{n}) + \mathbf{a}_{..}].$$

Cu (163) putem da expresii mai bune pentru unele matrici definite anterior. Elementele matricei $\mathbf{A}_n$ din § 3.3.2 sînt funcţii liniare de elementele x_{ij} şi putem să scriem

$$\mathbf{A}_n \doteq \mathbf{B}_n\mathbf{X}$$

unde $\mathbf{B}_n$ este o matrice simetrică $n \times n$ conţinînd ponderile funcţiei de dispersie. Mai precis, definim elementul (i, j) al matricei $\mathbf{B}_n$ ca $-b_{ij}$ dacă $i > j$ şi $-b_{ji}$ dacă $i > j$. Elementul diagonal i al matricei $\mathbf{B}_n$ este $b_i = \sum_{j \neq i} b_{ij}$. Astfel $\mathbf{B}_n$ are pentru elementele nediagonale valorile negative ale ponderilor funcţiei de dispersie şi elementele diagonale pozitive astfel că suma liniilor este zero.

Avem

$$\mathbf{V}_n = \mathbf{A}_n'\mathbf{A}_n = \mathbf{X}'\mathbf{B}_n\mathbf{B}_n\mathbf{X}$$

Din (163), matricea $\mathbf{C}_n$ din § 3.3.3 va avea elementul (k,l)

$$c_{kl} = -\left(\int f^2\right) \sum_{i<j} b_{ij}(x_{jk} - x_{ik})(x_{jl} - x_{il}).$$

Urmează că

$$\mathbf{C}_n = -\left(\int f^2\right) \mathbf{X}'\mathbf{B}_n\mathbf{X}.$$

Am exprimat direct matricile $\mathbf{V}_n$ şi $\mathbf{C}_n$ în funcţie de matricea $\mathbf{X}$ şi matricea de pondere $\mathbf{B}_n$. Deoarece sumele liniilor (şi coloanelor) lui $\mathbf{B}_n$ sînt zero, poate fi folosită matricea centrată $\mathbf{X}_c$ cu

$$\mathbf{V}_n = \mathbf{X}_c'\mathbf{B}_n\mathbf{B}_n\mathbf{X}_c$$

$$C_n = -\left(\int f^2\right)\mathbf{X}_c'\mathbf{B}_n\mathbf{X}_c.$$

Presupunerea (A8). Dacă are loc (163),

$$(1/n^2)\mathbf{C}_n \to \mathbf{C} \quad \text{cînd } n \to \infty,$$

unde $\mathbf{C}$ este o matrice $p \times p$ de rang maxim.

Acum definim o funcţie pătratică de Δ pentru a o folosi ca aproximaţie a lui $D^*(\Delta)$ prin

$$Q(\Delta) = -\Delta'\mathbf{C}\Delta + n^{-3/2}[\mathbf{a}_{..} - 2\mathbf{T}(\mathbf{0})]'\Delta + D(\mathbf{0}).$$

Scanned with OKEN Scanner

Atunci

$$\frac{\partial Q(\Delta)}{\partial \Delta} = -2\mathbf{C}\Delta + n^{-3/2}[\mathbf{a}_{..} - 2\mathbf{T}(\mathbf{0})].$$

LEMA 10. *Dacă are loc* (A8) *şi presupunerile teoremei 12 atunci*

$$\sup_{\Delta \in \mathcal{D}} n^{-3/2} \| \mathbf{T}(\Delta/\sqrt{n}) - \mathbf{T}(\mathbf{0}) - n^{-3/2}\mathbf{C}\Delta \| \xrightarrow{P} 0 \quad \textit{cînd } n \to \infty.$$

Demonstraţie. În expresia lui $\mathbf{R}(\Delta)$ se înlocuieşte $(1/n^2)\mathbf{C}_n$ prin $\mathbf{C}$.

LEMA 11. *Dacă au loc presupunerile lemei 10, atunci*

$$\sup_{\Delta \in \mathcal{D}} \left\| \frac{\partial Q(\Delta)}{\partial \Delta} - \frac{\partial D^*(\Delta)}{\partial \Delta} \right\| \xrightarrow{P} 0 \quad \textit{cînd } n \to \infty.$$

Demonstraţie. Notăm că

$$\frac{\partial Q(\Delta)}{\partial \Delta} - \frac{\partial D^*(\Delta)}{\partial \Delta} = 2n^{-3/2}[\mathbf{T}(\Delta/\sqrt{n}) - \mathbf{T}(\mathbf{0}) - n^{3/2}\mathbf{C}\Delta]$$

şi folosim lema 9.

LEMA 12. *Dacă au loc presupunerile lemei 10, atunci*

$$\sup_{\Delta \in \mathcal{D}} | Q(\Delta) - D^*(\Delta) | \xrightarrow{P} 0 \quad \textit{cînd } n \to \infty.$$

Egalînd derivatele parţiale cu zero, $Q(\Delta)$ are valoarea minimă în punctul

$$\Delta_n^* = n^{-3/2}\mathbf{C}^{-1}[(1/2)\mathbf{a} - \mathbf{T}(\mathbf{0})].$$

TEOREMA 13 (Sievers (1979)). *Dacă* (A8), *presupunerile teoremei 12 şi* (A5) *cu* $\gamma_n = n^{-3}$ *au loc atunci* $\hat{\Delta}_n$ *(definit de* Jaeckel (1972) *şi prezentat în* § 3.2) *şi* Δ_n^* *sînt asimptotic echivalenţi.*

Dacă au loc teoremele 11 şi 13, cînd $\boldsymbol{\beta} = \mathbf{0}$,

$$n^{-3/2}[\mathbf{T}(\mathbf{0}) - (1/2)\mathbf{a}_{..}] \xrightarrow[n\to\infty]{\mathcal{R}} N(\mathbf{0}, (1/12)\mathbf{V}).$$

Urmează că, asimptotic, $\Delta_n^* \sim N(\mathbf{0}, (1/12)\mathbf{C}^{-1}\mathbf{V}\mathbf{C}^{-1})$ şi avem

COROLARUL 5. *In presupunerile teoremelor 11 şi 13 cînd* $\boldsymbol{\beta}=\mathbf{0}$, *asimptotic*

$$\sqrt{n}\hat{\boldsymbol{\beta}}_n = \hat{\Delta}_n \sim N(\mathbf{0}, (1/12)\mathbf{C}^{-1}\mathbf{V}\mathbf{C}^{-1}).$$

Din invarianţa la translaţie a estimaţiei

$$\sqrt{n}\,(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})\,|_{\boldsymbol{\beta}} \xrightarrow{\mathcal{R}} \sqrt{n}\,\hat{\boldsymbol{\beta}}_n|_{\mathbf{0}}$$

Dacă $n\mathbf{C}_n^{-1}\mathbf{V}_n\mathbf{C}_n^{-1}$ este folosită ca o aproximaţie a lui $\mathbf{C}^{-1}\mathbf{V}\mathbf{C}^{-1}$, matricea de covarianţe asimptotică pentru $\hat{\boldsymbol{\beta}}_n$ este aproximativ

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}_n) = (1/12)\mathbf{C}_n^{-1}\mathbf{V}_n\mathbf{C}_n^{-1} =$$

$$= \left(1/12\left(\int f^2\right)^2\right)(\mathbf{X}'\mathbf{B}_n\mathbf{X})^{-1}(\mathbf{X}'\mathbf{B}_n\mathbf{B}_n\mathbf{X})(\mathbf{X}'\mathbf{B}_n\mathbf{X})^{-1}. \tag{164}$$

Scanned with OKEN Scanner

Observații. i) $\mathbf{X}_c$ poate fi folosită în locul matricei $\mathbf{X}$ în (164).

ii) (164) poate fi folosită în examinarea efectului diferitelor matrici de ponderare asupra estimației. Pentru cazul $b_{ij} \equiv 1$, cazul neponderat

$$\mathbf{B}_n = n(\mathbf{I}_n - (1/n)\mathbf{J}_n) \text{ și}$$

$$\text{cov}\,(\hat{\boldsymbol{\beta}}_n) = \left(1/12\left(\int f^2\right)^2\right)(\mathbf{X}_c'\mathbf{X}_c)^{-1}. \tag{165}$$

Deoarece

$$(\mathbf{X}'\mathbf{B}_n\mathbf{X})(\mathbf{X}'\mathbf{B}_n\mathbf{B}_n\mathbf{X})^{-1}(\mathbf{X}'\mathbf{B}_n\mathbf{X})^{-1} = [\mathbf{X}'\mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{X}]^{-1},$$

cov $(\hat{\boldsymbol{\beta}}_n)$ depinde de $\mathbf{B}_n$ împreună cu matricea $\mathbf{H} = \mathbf{B}_n\mathbf{X}$. Matricea $\mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'$ este matricea proiecție, proiecția fiind în spațiul coloană al lui $\mathbf{H}$.

Observații. i) În cazul neponderat, $b_{ij} \equiv 1$ și $\mathbf{B}_n = n(\mathbf{I}_n - (1/n)\mathbf{J}_n)$ Acest caz dă o eficiență asimptotică înaltă, după cum se poate vedea exprimînd diferența matricilor de covarianțe (164) și (164) ca (exceptînd o constantă de multiplicare)

$$(\mathbf{X}_c'\mathbf{B}_n\mathbf{X}_c)^{-1}(\mathbf{X}_c'\mathbf{B}_n\mathbf{B}_n\mathbf{X}_c)(\mathbf{X}_c'\mathbf{B}_n\mathbf{X}_c)^{-1} - (\mathbf{X}_c'\mathbf{X}_c)^{-1} = \mathbf{M}\mathbf{M}',$$

unde $\mathbf{M} = (\mathbf{X}_c'\mathbf{B}_n\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{B}_n - (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'$. Matricea $\mathbf{M}\mathbf{M}'$ este semipozitiv definită și urma și determinantul lui (164) nu pot fi mai mici decît cele ale lui (165).

(164) și (165) sînt egale dacă și numai dacă $\mathbf{M} = \mathbf{0}$ sau

$$(\mathbf{X}_c'\mathbf{B}_n\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{B}_n = (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'.$$

Echivalent

$$\mathbf{B}_n\mathbf{X}_c = \mathbf{X}_c(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{B}_n\mathbf{X}_c.$$

Deoarece $\mathbf{X}_c(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'$ este proiecția în spațiul coloană a lui $\mathbf{X}_c$, egalitatea va avea loc dacă și numai dacă, coloanele matricei $\mathbf{B}_n\mathbf{X}_c$ sînt în spațiul coloană al matricei $\mathbf{X}_c$($\mathbf{B}_n\mathbf{X}_c = \mathbf{X}_c\mathbf{G}$, pentru o matrice $\mathbf{G} : p \times p$).

ii) Folosirea ponderilor $b_{ij} \not\equiv 1$ nu poate crește eficiența estimației față de cazul neponderat. Este posibil să alegem ponderile astfel încît să avem o pierdere mică de eficiență și să avem cîștig în raport cu alte aspecte.

iii) Folosind ponderile $b_{ij} = 0$ sau 1 putem reduce numărul de termeni din funcția de dispersie D, reducînd numărul de comparări. Aceasta poate avea avantaje de calcul. Exemplele pe care le vom da arată că nu avem pierdere de eficiență. Cu astfel de alegere a ponderilor funcția de dispersie depinde de rangurile *restrînse* ale rezidurilor Z_i.

Folosind (158′),

$$D = \sum_{i=1}^{n} (2R_i - (\#\ C_i + 1))\, Z_i,$$

unde $C_i = \{j : b_{ij} = 1,\ 1 \leqslant j \leqslant n,\ j \neq i\}$, numărul de elemente din C_i este notat prin $\#\ C_i$ și R_i este rangul lui Z_i în mulțimea $\{Z_j : j \in C_i\}$.

Ponderile b_{ij} sînt asociate direct cu observațiile (nu cu observațiile ordonate) și pot fi alese să depindă de $\mathbf{X}$.

Scanned with OKEN Scanner

Cînd $\mathbf{X}_c'\mathbf{X}_c$ este aproape singulară, folosirea ponderilor poate reduce efectele multicolinearității.

Un mod de luare a ponderilor este de a atribui o pondere w_i pentru fiecare observație $i = 1, \ldots, n$ și să folosim $b_{ij} = w_i w_j$, $i \neq j$. Pentru simplificare presupunem că $\sum_{i=1}^{n} w_i = 1$. Definim o matrice $\mathbf{W}$ $n \times n$ diagonală care are ca element diagonal i pe w_i și vectorul $w = (w_1, \ldots, w_n)'$. Atunci matricea de pondere este

$$\mathbf{B}_n = \mathbf{W} - \mathbf{ww}' = (\mathbf{I}_n - \mathbf{WJ}_n)\mathbf{W}(\mathbf{I}_n - \mathbf{J}_n\mathbf{W}) =$$

$$= (\mathbf{I}_n - \mathbf{WJ}_n)(\mathbf{W} - \mathbf{ww}')(\mathbf{I}_n - \mathbf{J}_n\mathbf{W}).$$

Notăm că $\mathbf{B}_n\mathbf{X}$ are ca element (i, j) pe $w_i(x_{ij} - \bar{x}_j^*)$, unde $\bar{x}_j^* = \sum_i w_i x_{ij}$ este media ponderată a coloanei j a lui $\mathbf{X}$. Atunci $\mathbf{B}_n\mathbf{X}$ este matricea ponderată și centrată.

Exemplul 1. Presupunem că avem n_1 observații (Grupul 1) urmînd modelul

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + e_i^*, \quad i = 1, \ldots, n_1$$

și un alt grup de n_2 observații urmînd modelul

$$Y_i = \beta_0^* + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + e_i^*, \quad i = n_1 + 1, \ldots, n_1 + n_2 = n.$$

Avem intercepții diferite și erorile pot avea repartiții deosebite. Ne propunem să estimăm $\beta_1, \ldots, \beta_p$. În unele situații nu pot fi comparate observațiile din diferite grupe. Grupele se pot referi la diferite variabile. În astfel de cazuri comparările dintre grupe pot fi excluse folosind

$$b_{ij} = \begin{cases} 1/n_1, & 1 \leqslant i < j \leqslant n_1 \\ 1/n_2, & n_1 + 1 \leqslant i < j \leqslant n_1 + n_2 \\ 0, & \text{în rest} \end{cases}$$

Atunci

$$D = (1/n_1) \sum_{i<j\leqslant n_1} |Z_i - Z_j| + (1/n_2) \sum_{n_1<i<j} |Z_i - Z_j|$$

și matricea de ponderi este

$$\mathbf{B}_n = \begin{pmatrix} \mathbf{I}_{n_1} - (1/n_1)\mathbf{J}_{n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_2} - (1/n_2)\mathbf{J}_{n_2} \end{pmatrix}.$$

Aceasta este o matrice idempotentă și (164) se reduce la

$$\text{cov}\,(\hat{\beta}_1, \ldots, \hat{\beta}_p) = \left(1/12 \left(\int f^2\right)^2\right)(\mathbf{X}'\mathbf{B}_n\mathbf{X})^{-1}. \tag{166}$$

Scanned with OKEN Scanner

Dacă toate erorile au aceeaşi repartiţie, definim variabila indicator pentru grupe

$$x_{i,\,p+1} = \begin{cases} 0, & 1 \leqslant i \leqslant n_1 \\ 1, & n_1 < i \leqslant n_1 + n_2 \end{cases}$$

şi adăugăm acest termen la model. Atunci folosim $b_{ij} \equiv 1$, funcţia de dispersie neponderată. Din (165) folosind o matrice $\mathbf{X}^*$ cu o coloană adiţională

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_p, \hat{\boldsymbol{\beta}}_{p+1}) = \left(1/12\left(\int f^2\right)^2\right)(\mathbf{X}_c^{*\prime}\mathbf{X}_c^*)^{-1}.$$

Cînd matricea de covarianţe a estimatorilor $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_p$ este determinată obişnuit, găsim că concordă cu (166). Astfel amîndouă metodele dau aceeaşi matrice de covarianţe asimptotică pentru estimatori. Omiterea comparărilor dintre grupe nu afectează eficienţa estimatorilor.

§ 4. Relaţii asimptotice ale estimatorilor M şi R în modelul liniar al regresiei

4.1. Presupuneri şi notaţii

1°. Pentru $N = 1, 2, \ldots,$ fie $Y_{N1}, \ldots, Y_{NN}$ variabile aleatoare independente astfel încît $Y_{Ni} = \sum_{j=1}^{p} \Delta_j c_{ji} + e_i$ are funcţia de repartiţie

$$F\left(y - \sum_{j=1}^{p} \Delta_j^0 c_{ji}\right), \quad i = 1, \ldots, N. \tag{167}$$

De asemenea, presupunem că $f(x) = \mathrm{d}F(x)/\mathrm{d}x$ există, este absolut continuă şi că are informaţia Fisher finită,
Fie $\varphi(t, f)$ definită ca în (119) cu

$$F^{-1}(t) = \inf\{y : F(y) \geqslant t\}.$$

2°. Fie $\mathbf{c}'_{(j)} = (c'_{j1}, \ldots, c'_{jN})$ şi $\mathbf{c}''_{(j)} = (c''_{j1}, \ldots, c''_{jN})$ vectori satisfăcînd

$$\text{(a)} \qquad \sum_{i=1}^{N} c'_{ji} = \sum_{i=1}^{N} c''_{ji} = 0, \quad j = 1, \ldots, p$$

şi

$$\text{(168)} \quad \mathbf{c}'_{(j)} \cdot (\mathbf{c}'_{(j)})^{\mathrm{T}*)} \leqslant M\,;\ \mathbf{c}''_{(j)} \cdot (\mathbf{c}''_{(j)})^{\mathrm{T}} \leqslant M,\ j = 1, \ldots, p\,;\ N = 1, 2, \ldots$$

($M > 0$ este o constantă independentă de N) unde fiecare din produsele scalare din (168) este fie 0 pentru toţi însă un număr finit de N sau pozitiv pentru toţi sau un număr finit de N; dacă $\mathbf{c}'_{(j)} \cdot (\mathbf{c}'_{(j)})^{\mathrm{T}} > 0$ pentru $N > N'$, presupunem

$$\text{(169)} \qquad \lim_{N\to\infty}\left\{\max_{1\leqslant i\leqslant n}(c'_{ji})^2\left[\sum_{k=1}^{N}(c'_{jk})^2\right]^{-1}\right\} = 0$$

*) în acest paragraf am notat transpusa prin $(\ )^{\mathrm{T}}$.

Scanned with OKEN Scanner

(condiția lui Noether) și o presupunere analoagă pentru $\mathbf{c}''_{(j)}$, $j = 1, \ldots, p$

(b) (A.3.c) din § 3.1.1.

Fie $\mathbf{C}_N = (c_{ji})_{\substack{1 \leq i \leq N \\ 1 \leq j \leq p}}$ matricea definită în §. 3.1.1.(A3) și care satisface (A.3.a), $\mathbf{c}'_{(j)}$ și $\mathbf{c}''_{(j)}$ satisfăcînd (a) și (b).

Mai mult, presupunem că

(c) $\lim_{N \to \infty} \mathbf{\Sigma}_N = \mathbf{\Sigma}$ există și $\mathbf{\Sigma}$ este o matrice pozitiv definită, unde

$$\mathbf{\Sigma}_N = \mathbf{C}_N \mathbf{C}_N^{\mathrm{T}}. \tag{170}$$

3°. Fie $\varphi(t)$, $0<t<1$, o funcție neconstantă care este exprimabilă ca o sumă finită de funcții monotone, de pătrat integrabile pe (0,1).

4°. Fie $\psi(x)$, $x \in R^1$ o funcție neconstantă care este exprimabilă ca o sumă finită de funcții monotone; presupunem că:

$$\int_{R^1} \psi^2(x) f(x)\, \mathrm{d}x < \infty. \tag{171}$$

Fie γ și A^2 definite ca în (120), respectiv (121),

$$\omega = -\int_{R^1} \psi(x) f'(x)\, \mathrm{d}x \tag{172}$$

și

$$\rho^2 = \int_{R^1} \psi^2(x) f(x)\, \mathrm{d}x - \left(\int_{R^1} \psi(x) f(x) \mathrm{d}x \right)^2. \tag{173}$$

Pentru orice vector $\Delta = (\Delta_1, \ldots, \Delta_p)$ fixat, fie

$$\delta_i(\Delta) = Y_{Ni} - \sum_{j=1}^{p} \Delta_j c_{ji}, \qquad i = 1, \ldots, N. \tag{174}$$

Fie $\hat{\Delta}_M$ estimația M a lui Δ^0 corespunzînd la funcția ψ, adică $\hat{\Delta}_M$ este soluția sistemului de ecuații

$$M_j(\mathbf{Y}_N, \Delta) = \sum_{i=1}^{N} c_{ji} \psi(\delta_i(\Delta)) = 0, \qquad j = 1, \ldots, p \tag{175}$$

în raport cu Δ.

Fie $\hat{\Delta}_R$ estimația rang corespunzătoare funcției φ, adică $\hat{\Delta}_R$ este orice soluție a problemei de minimizare

$$\sum_{j=1}^{p} \left| \sum_{i=1}^{N} c_{ji} a_N(R_i^{\Delta}) \right| = \min., \tag{176}$$

unde R_i^{Δ} este rangul lui $\delta_i(\Delta)$ printre $\delta_1(\Delta), \ldots, \delta_N(\Delta)$;

$$R_i^{\Delta} = \sum_{j=1}^{N} u(\delta_i(\Delta) - \delta_j(\Delta)), \tag{177}$$

unde

$$u(x) = \begin{cases} 1, & x \geqslant 0 \\ 0, & x < 0, \end{cases}$$

a_N este funcția scor corespunzătoare lui φ definită ca în (118).

Scanned with OKEN Scanner

4.2. Repartiția asimptotică a statisticii $\hat{\Delta}_M - \hat{\Delta}_R$

TEOREMA 14 (Jurečková (1977)). *In presupunerile 1°—4° și sub* $\gamma \neq 0$, $\omega \neq 0$, $\hat{\Delta}_M - \hat{\Delta}_R$ *are pentru* $N \to \infty$, *o repartiție normală cu vectorul valoare medie* **0** *și matricea de covariație*

$$\Sigma^{-1} \cdot \int_0^1 \left[\frac{1}{\omega} (\psi(F^{-1}(t)) - \bar{\psi}) - \frac{1}{\gamma} (\varphi(t) - \bar{\varphi}) \right]^2 \mathrm{d}t, \tag{178}$$

unde $\bar{\psi} = \int \psi(x) \mathrm{d}F(x)$, $\bar{\varphi} = \int_0^1 \varphi(t) \mathrm{d}t$.

COROLARUL 6 (Jurečková (1977)). *In presupunerile 1°—4° și sub* $\gamma \neq 0$, $\omega \neq 0$. $\hat{\Delta}_M$ *este asimptotic echivalent în probabilitate cu* $\hat{\Delta}_R$ *(în sensul că* $\| \hat{\Delta}_M - \hat{\Delta}_R \| \xrightarrow{P} 0$ *cînd* $N \to \infty$*) dacă și numai dacă*

$$\varphi(t) = a\psi(F^{-1}(t)) + b, \quad t \in (0, 1) \tag{179}$$

pentru unii $a > 0$, $b \in R^1$.

Urmează ca estimațiile M și R au asimptotic o proprietate minimax pe mulțimea estimațiilor asimptotic nedeplasate în modelul de contaminare simetrică.

Punem

$$\psi(x) = -\frac{g'(x)}{g(x)}, \quad x \in R^1 \tag{180}$$

unde g este o densitate unimodală cu informația Fisher finită și astfel ca

$$\int_{R^1} (g'(x)/g(x))^2 f(x) \mathrm{d}x < \infty.$$

Atunci $\hat{\Delta}_M$ este estimația de verosimilitate maximă corespunzătoare lui g. Similar, punem

$$\varphi(t) = \varphi(t, g), \quad 0 < t < 1 \tag{181}$$

Atunci $\hat{\Delta}_R$ corespunzător la φ este estimația rang, asimptotic eficient pentru g în rolul repartiției de bază. Din teorema 14 urmează că repartiția asimptotică a statisticii $\hat{\Delta}_M - \hat{\Delta}_R$ este atunci normală cu centrul **0** și matricea de covariație

$$\Sigma^{-1} \int_0^1 \left[\frac{1}{\omega} \left(\frac{g'(F^{-1}(t))}{g(F^{-1}(t))} - \bar{g} \right) + \frac{1}{\gamma} \varphi(t, g) \right]^2 \mathrm{d}t, \tag{182}$$

unde

$$\bar{g} = \int_{R^1} \frac{g'(x)}{g(x)} \mathrm{d}F(x).$$

Sub (180) și (181), avem următoarele corolare:

Scanned with OKEN Scanner

COROLARUL 7 (Jurečková (1977). *Fie* ψ *şi* φ *satisfăcînd* (180) *şi respectiv* (181), *cînd* g *este densitatea repartiţiei* $N(0, \sigma^2)$, $\sigma^2 > 0$. *Atunci* $\hat{\Delta}_M$ *este asimptotic echivalent în probabilitate cu* $\hat{\Delta}_R$ (*se notează* $\hat{\Delta}_M \sim \hat{\Delta}_R$) *dacă şi numai dacă repartiţia de bază* f *este normală* $N(a, \lambda^2)$ *pentru* $\lambda^2 > 0$, $a \in R^1$.

COROLARUL 8 (Jurečková (1977)). *Fie* ψ *şi* φ *satisfăcînd* (180) *şi respectiv* (181) *cu* g *densitatea logistică. Atunci* $\hat{\Delta}_M \sim \hat{\Delta}_R$ *dacă şi numai dacă* $f \equiv g$.

Observaţie. Fie ψ şi φ satisfăcînd (180) şi respectiv (181) cu g densitatea repartiţiei dublu-exponenţială, $g(x) = \frac{1}{2} e^{-|x|}$, $x \in R^1$. Atunci $\hat{\Delta}_M$ şi $\hat{\Delta}_R$ coincid pentru orice repartiţie f simetrică.

4.3. Comportarea asimptotică a $M_j(\mathbf{Y}_N, \Delta)$

În demonstraţia teoremei 14 se utilizează aproximaţia lui $M_j(\mathbf{Y}_N, \Delta)$ printr-o funcţie liniară de Δ.

TEOREMA 15 (Jurečková (1977)). *In presupunerile 1°, 2°, şi 4°, pentru orice* $K > 0$, $\varepsilon > 0$ *şi* $j = 1, \ldots, p$ *are loc*

$$\lim_{N\to\infty} P_{\Delta^0}\{\max_{\|\Delta-\Delta^0\|<K} |M_j(\mathbf{Y}_N, \Delta) - M_j(\mathbf{Y}_N, \Delta^0) + \omega(\Delta - \Delta^0)\sigma^{(j)}| \geqslant \varepsilon\} = 0, \tag{183}$$

unde $\sigma^{(j)}$ *este coloana* j *a lui* Σ.

COROLARUL 9. *Dacă* $\{\Delta_N^*\}_{1<N<\infty}$ *este un şir de vectori aleatori astfel încît* $\Delta_N^* - \Delta^0$ *ete mărginit în probabilitate şi dacă sînt satisfăcute 1°, 2° şi 4°, atunci*

$$\lim_{N\to\infty} P_{\Delta^0}\{|M_j(\mathbf{Y}_N, \Delta_N^*) - M_j(\mathbf{Y}_N, \Delta^0) + \omega(\Delta_N^* - \Delta^0)\sigma^{(j)}| \geqslant \varepsilon\} = 0 \tag{184}$$

pentru $j = 1, \ldots, p$ *şi orice* $\varepsilon > 0$.

Demonstraţia teoremei 14 este bazată pe aplicarea corolarului 9 la $\Delta_N^* = \hat{\Delta}_M$, astfel că trebuie să cunoaştem dacă $(\hat{\Delta}_M - \Delta^0)$ este mărginit în probabilitate. Următoarea lemă arată că acesta este cazul sub presupunerile 1°, 2° şi 4°.

LEMA 13. *In presupunerile 1°, 2° şi 4°, la orice* $\varepsilon > 0$ *corespund* $K > 0$, $\eta > 0$ *şi un întreg pozitiv* N_0 *astfel încît pentru* $N > N_0$ *are loc*

$$P_{\Delta^0}\{\max_{\|\Delta-\Delta^0\|>K} \|M(\mathbf{Y}_N, \Delta)\| < \eta\} < \varepsilon, \tag{185}$$

unde

$$M(\mathbf{Y}_N, \Delta) = (M_1(\mathbf{Y}_N, \Delta), \ldots, M_p(\mathbf{Y}_N, \Delta)). \tag{186}$$

§ 5. Asupra estimării densităţii de repartiţie a erorilor în regresie

5.1. Introducere

În teoria generală a regresiei se presupune că erorile sînt variabile aleatoare independente identic repartizate $N(0, \sigma^2)$, σ^2 finită. În multe cazuri această presupunere este suspectă. Cercetătorul poate fi interesat

Scanned with OKEN Scanner

în cunoaşterea formei densităţii de repartiţie a erorilor. În particular, prezintă importanţă să găsim dacă repartiţia erorilor este o mixtură de repartiţii, fapt ce poate explica faptul că densitatea erorilor este bimodală. În aceste condiţii prezintă interes estimarea densităţii de repartiţie a erorilor. Fazal (1977) estimează densitatea de repartiţie a erorilor a cărei valoare medie este presupusă a fi zero. Face acest lucru folosind estimaţii de tip nucleu ale densităţilor, estimaţii propuse de Rosenblatt (1956) şi studiate printre alţii de Parzen (1962) şi Cacoullos (1968).

Dăm mai întîi unele rezultate ale lui Cacoullos (1968) care generalizează rezultatele lui Parzen (1962).

TEOREMA 16. *Fie* $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ *vectori aleatori q-dimensionali independenţi şi identic repartizaţi avînd funcţia (densitatea) de repartiţie* $F(f)$. *Presupunem că* W *este o funcţie de pondere simetrică definită pe* R^q *ce satisface condiţiile*

$$\int W(\mathbf{u})\mathrm{d}\mathbf{u} = 1, \quad \sup_{\mathbf{u}} |W(\mathbf{u})| < \infty, \quad \lim_{\|\mathbf{u}\|\to\infty} \|\mathbf{u}\|^q \, |W(\mathbf{u})| = 0,$$

$$\int |W(\mathbf{u})|^{2+\delta}\mathrm{d}\mathbf{u} < \infty$$

pentru δ *pozitiv mic. Aici* $\|\mathbf{u}\|^2 = \sum_{i=1}^{q} u_i^2$.

Fie $\{b_n\}$ *un şir de numere pozitive astfel încît*

$$b_n = o(1) \quad \text{şi} \quad nb_n^q = o(1).$$

Fie

$$\mathbf{u}/b_n = (u_1/b_n, \ldots, u_q/b_n) \quad \text{şi} \quad W(\mathbf{u}) = b_n^{-q} W(\mathbf{u}/b_n).$$

Fie $F_n(\mathbf{v})$ *funcţia de repartiţie empirică. Definim o estimaţie a densităţii prin*

$$f_n(\mathbf{u}) = \int W_n(\mathbf{u} - \mathbf{v})\,\mathrm{d}F_n(\mathbf{v}) = \frac{1}{n}\sum_{i=1}^{n} W_n(\mathbf{u} - \mathbf{Z}_i). \tag{187}$$

a) *Dacă* f *are derivatele parţiale de ordinul trei mărginite şi* $\mathbf{u}$ *este un punct de continuitate al funcţiei* f, *atunci estimaţia* $f_n(\mathbf{u})$ *este consistentă în medie patratică. Mai mult, dacă* $nb_n^{q+4} = o(1)$, *atunci*

$$(nb_n)^{-q/2}(f_n(\mathbf{u}) - f(\mathbf{u})) \xrightarrow{\mathcal{R}} N(0, f(\mathbf{u})\int W^2(\mathbf{v})\mathrm{d}\mathbf{v}).$$

b) *Dacă* f *este uniform continuă,* $nb_n^{2q} = o(1)$ *şi transformata Fourier a funcţiei* $W(\mathbf{u})$ *este absolut integrabilă atunci* $f_n(\mathbf{u})$ *este consistentă uniform în probabilitate, adică*

$$\sup_{\mathbf{u}} |f_n(\mathbf{u}) - f(\mathbf{u})| \xrightarrow{P} 0.$$

Consistenţa uniformă a funcţiei $f_n(\mathbf{u})$ cu probabilitatea unu a fost studiată de Nadaraya (1965) şi Schuster (1969). Rezultatele lor sînt rezumate în cazul multidimensional de următoarea teoremă

Scanned with OKEN Scanner

Scanned with OKEN Scanner

TEOREMA 17. *Fie $Z_1, \ldots, Z_n$ vectori q-dimensionali independenți și identic repartizați cu funcția de repartiție F care are o densitate uniform continuă. Definim o estimație a densității de repartiție f prin* (187) *unde funcția de pondere W este ne-negativă, este de variație mărginită și* $\int W(\mathbf{v})\,d\mathbf{v} = 1$.

Presupunem că pentru fiecare pozitiv real γ, seria

$$\sum_n \exp\{-\gamma n b_n^{2q} \alpha_n^{-2}\} \text{ converge.}$$

Șirul de numere nenegative $\{\alpha_n\}$ și $\{b_n\}$ este astfel

$$\alpha_n^{-1} = o(1), \quad \alpha_n b_n = o(1) \text{ și } n^{-1} b_n^{-2} \alpha_n^2 = o(1);$$

atunci

$$\sup_{\mathbf{u}} \alpha_n |f_n(\mathbf{u}) - f(\mathbf{u})| \xrightarrow{\text{a.s.}} 0.$$

5.2. O estimație a densității de repartiție a erorii

Presupunem că avem n perechi de observații independente (y_i, x_{in}) $i = 1, \ldots, n$ din modelul de regresie

$$y_i = \alpha + \beta x_{in} + \varepsilon_i, \tag{188}$$

unde ε_i sînt variabile aleatoare independente identic repartizate cu densitatea de repartiție g necunoscută cu media 0 și dispersia necunoscută σ^2 finită.

Presupunem că $\sum_{i=1}^{n} x_{in} = 0$. Definim

$$z_{in} = x_{in} / (\sum_i x_{in}^2)^{1/2}. \tag{189}$$

Predictorul de cele mai mici pătrate pentru modelul (188) este

$$\hat{y}_i = \bar{y} + (\sum z_{jn} y_j) z_{in}$$

unde

$$\bar{y} = \frac{1}{n} \sum_i y_i.$$

Fie $\bar{\varepsilon} = \frac{1}{n} \sum \varepsilon_i$ (Dacă nu se specifică, suma se ia de la 1 pînă la n).

O estimație a erorii este rezidul de cele mai mici pătrate notat prin $\hat{\varepsilon}_i = y_i - \hat{y}_i$. Avem

$$\varepsilon_i - \hat{\varepsilon}_i = \bar{\varepsilon} + z_{in} \sum_j z_{jn} \varepsilon_j. \tag{190}$$

Fie

$$g_n(\mathbf{u}) = \frac{1}{n} \sum_i W_n(\mathbf{u} - \varepsilon_i). \tag{191}$$

O estimație a densității de repartiție $g(\mathbf{u})$ este :

$$\hat{g}(\mathbf{u}) = \frac{1}{n} \sum_i W_n(\mathbf{u} - \hat{\varepsilon}_i). \tag{192}$$

Funcțiile W și W_n și șirul b_n sînt descrise în teorema 16 (cu $q = 1$). După teorema 16.a),

$$g_n(u) = g(u) + \mathbf{0}_P((nb_n)^{-1/2}). \tag{193}$$

Arătam că în presupuneri suplimentare, efectul folosirii variabilelor estimator $\hat{\varepsilon}_i$ în (192) este neglijabil.

TEOREMA 18 (Fazal (1977)). *Fie $g_n(u)$ și $\hat{g}_n(u)$ definite ca mai înainte. Presupunem că funcția de pondere W satisface următoarele condiții suplimentare*

i) *W este Lipschitziană de ordinul $\lambda \in (2/3, 1]$, pentru numărul pozitiv mic δ*

ii) $\lim_n \sum_j | x_{jn}^2 / \sum_i x_{in}^2 |^{1+\delta/2} = 0$ *și*

iii) $M(|\varepsilon_1|^{2+\delta}) < \infty$.

Atunci

a) $\sup_u |\hat{g}_n(u) - g_n(u)| = \mathbf{0}_P(n^{-\lambda/2} b_n^{-1-\lambda})$.

b) $|\hat{g}_n(u) - g_n(u)| = \mathbf{0}_P(n^{-\lambda/2} b_n^{-1-\lambda})$.

c) *Dacă condițiile date de partea b) a teoremei 16 sînt de asemenea satisfăcute, atunci*

$$\sup_u |\hat{g}_n(u) - g(u)| = \mathbf{0}_P(n^{-\lambda/2} b_n^{-1-\lambda}).$$

Demonstrație. Fie

$$R_n(u) = \hat{g}_n(u) - g_n(u), \tag{194}$$

adică

$$R_n(u) = (nb_n)^{-1} \sum_i [W\{(u - \hat{\varepsilon}_i)/b_n\} - W\{(u - \varepsilon_i)/b_n\}].$$

După condiția Lipschitz asupra lui W, avem pentru unele constante pozitive C_λ și $\lambda \in (2/3, 1]$,

$$|R_n(u)| \leqslant C_\lambda n^{-1} b_n^{-1-\lambda} \sum_i |\hat{\varepsilon}_i - \varepsilon_i|^\lambda \leqslant$$

$$\leqslant C_\lambda b_n^{-1-\lambda} (|\bar{\varepsilon}|^\lambda + n^{-1} \sum_i |z_{in} \sum_j z_{jn} \varepsilon_j|^\lambda). \tag{195}$$

Scriem (195) ca $A_n + B_n$, unde

$$A_n = C_\lambda n^{-\lambda/2} b_n^{-1-\lambda} |n^{1/2} \bar{\varepsilon}|^\lambda. \tag{196}$$

Scanned with OKEN Scanner

Din teorema limită centrală pentru variabile aleatoare independente, identic repartizate, urmează că

$$A_n = \mathbf{0}_p(n^{-\lambda/2} b_n^{-1-\lambda}),$$

$$B_n = C_\lambda n^{-1} b_n^{-1-\lambda} \Big| \sum_i z_{in} \sum_j z_{jn} \varepsilon_j \Big|^\lambda \leqslant$$

$$\leqslant C_\lambda n^{-1} b_n^{-1-\lambda} \sum_i |z_{in}|^\lambda \Big| \sum_j z_{jn} \varepsilon_j \Big|^\lambda \leqslant$$

$$\leqslant C_\lambda n^{-\lambda/2} b_n^{-1-\lambda} \Big| \sum_i z_{in}^2 \Big|^{\lambda/2} \Big| \sum_{j_i} z_{jn} \varepsilon_j \Big|^\lambda \leqslant$$

$$\leqslant C_\lambda n^{-\lambda/2} b_n^{-1-\lambda} \Big| \sum_j z_{in} \varepsilon_j \Big|^\lambda \text{ (după (189)).}$$

După presupunerile ii) și iii) sînt satisfăcute condițiile lui Liapunov pentru șirul de variabile aleatoare independente $\{z_{in}\, \varepsilon_i\}$.

Deci

$$B_n = \mathbf{0}_p(n^{-\lambda/2} b_n^{-1-\lambda}).$$

Aceasta împreună cu (194) și (196) dau

$$\sup_u |\hat{g}_n(u) - g_n(u)| = \mathbf{0}_p(n^{-\lambda/2} b_n^{-1-\lambda}). \tag{197}$$

Pentru a arăta b) considerăm

$$|\hat{g}_n(u) - g(u)| \leqslant \sup_u |\hat{g}_n(u) - g_n(u)| + |g_n(u) - g(u)|.$$

Al doilea termen din membrul drept al inegalității după teorema 16.a) este $\mathbf{0}_p(b_n^{-1/2} n^{-1/2})$. Aceasta și (197) dau al doilea rezultat.

Pentru partea c) folosim

$$\sup_u |\hat{g}_n(u) - g(u)| \leqslant \sup_u |\hat{g}_n(u) - g_n(u)| + \sup_u |g_n(u) - g(u)|.$$

Teorema 16.b) implică că:

$$\sup_u |g_n(u) - g(u) = o_p(1).$$

Deoarece una dintre condițiile teoremei 16.a) este că $nb_n^5 = o(1)$ și deoarece din cele anterioare trebuie să ne restrîngem la $nb_n^{2(1+1/\lambda)} \to \infty$ cînd $n \to \infty$ este necesar ca λ să varieze în intervalul $(2/3, 1]$.

5.3. Cazul uniform (rectangular) al funcției de pondere

Condițiile teoremei 18 nu vor avea loc pentru funcția de pondere rectangulară $W(u) = \frac{1}{2} a\, I_{[-a,a]}(u)$, unde a este un număr real nenegativ, și

$$I_A(u) = \begin{cases} 1, & u \in A \\ 0, & \text{în rest.} \end{cases}$$

Scanned with OKEN Scanner

Fie $a = 1$. În acest caz $g_n(u)$ și $\hat{g}_n(u)$ sînt date de

$$g_n(u) = (2nb_n)^{-1} \sum_i I_{[-b_n, b_n]}(u - \varepsilon_i) \tag{198}$$

și

$$\hat{g}_n(u) = (2nb_n)^{-1} \sum_i I_{[-b_n, b_n]}(u - \hat{\varepsilon}_i). \tag{199}$$

TEOREMA 19 (Brillinger (1969), Rosenkrantz și O'Reilly (1972)). *Fie $Z_1, \ldots, Z_n$ variabile aleatoare independente identic repartizate avînd funcția de repartiție F. Fie F_n funcția de repartiție empirică. Atunci*

$$\sup_u |F_n(u) - F(u)| = \mathbf{0}_p(n^{-1/2}) + \mathbf{0}_p(n^{-3/4})(\log n)^{1/2}(\log\log n)^{1/4}).$$

TEOREMA 20 (Fazal (1977)). *Fie șirul de parametrii b_n ce satisfac condițiile teoremei* 16. *Fie $g_n(u)$ și $\hat{g}_n(u)$ definite de* (198) *și* (199). *Dacă densitatea g este mărginită și $(nb_n^2)^{-1} = o(1)$, atunci*

a) $$\sup_u |\hat{g}_n(u) - g_n(u)| = \mathbf{0}_p(n^{-1/2} b_n^{-1}) +$$

$$+ \mathbf{0}_p(n^{-3/4}(\log n)^{1/2}(\log\log n)^{1/4} b_n^{-1}) +$$

$$+ \mathbf{0}_p(\max_{1 \leqslant i \leqslant n} |z_{in}| b_n^{-1}).$$

Membrul drept al ultimei egalități va fi notat prin D_n. Atunci

b) $|\hat{g}_n(u) - g(u)| = D_n$.

Demonstrație. Fie „$\# \varepsilon_j = c$" numărul ε_j—lor egali cu c. Fie G_n funcția de repartiție empirică a lui ε_j, $j = 1, \ldots, n$, adică

$$G_n(u) = (\# \varepsilon_j \leqslant u)/n.$$

Atunci (198) poate fi scris ca

$$g_n(u) = (2b_n)^{-1}[G_n(u + b_n) - G_n(u - b_n)].$$

Relația (190) implică că

$$\varepsilon_i - \hat{\varepsilon}_i = \mathbf{0}_p(n^{-1/2}) + \mathbf{0}_p(|z_{in}|) = -\rho_n \text{ (notație)}. \tag{200}$$

Atunci (199) poate fi scris ca

$$\hat{g}_n(u) = (2b_n)^{-1}[G_n(u + b_n - \rho_n) - G_n(u - b_n - \rho_n)].$$

Din teorema 19 deducem că

$$\sup_u |\hat{g}_n(u) - g_n(u)| \leqslant (2b_n)^{-1} \sup_u |[G(u + b_n - \rho_n) - G(u + b_n)] -$$

$$- [G(u - b_n - \rho_n) - G(u - b)]| + \mathbf{0}_p(n^{-1/2} b_n^{-1}) + \tag{201}$$

$$+ \mathbf{0}_p(n^{-3/4}(\log n)^{1/2}(\log\log n)^{1/4} b_n^{-1}).$$

Scanned with OKEN Scanner

Dacă densitatea g este mărginită, atunci

$$b_n^{-1} \sup_u |G(u \pm b_n - \rho_n) - G(u \pm b_n)| = \mathbf{0}_p(\rho_n b_n^{-1}) =$$

$$= \mathbf{0}_p(n^{-1/2} b_n^{-1}) + \mathbf{0}_p(\max_{1 \leq i \leq n} |z_{in}| b_n^{-1}) \text{ după (200).}$$

Această ultimă relație și (201) demonstrează partea a). Al doilea rezultat este implicat de (193).

5.4. O extensie la regresia multiplă

Fie

$$\mathbf{y} = \alpha \mathbf{1}_n + \mathbf{X}_n \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

unde: $\mathbf{y}$ este vectorul $(n \times 1)$ al variabilelor dependente,
$\mathbf{1}_n' = (1, \ldots, 1)$
$\mathbf{X}_n$ este matricea $(n \times r)$ de constante (variabile independente de rang r (r este fixat).
(Fără a pierde din generalitate presupunem că $\mathbf{1}_n' \mathbf{X}_n = \mathbf{0}$)

$$\boldsymbol{\varepsilon}' = (\varepsilon_1, \ldots, \varepsilon_n)$$

este vectorul $(1 \times n)$ al erorilor (erorile ε_i satisfac aceleași condiții ca în cazul regresiei simple.

Predictorii de cele mai mici pătrte sînt

$$\hat{\mathbf{y}} = \bar{y} \mathbf{1}_n + \mathbf{X}_n (\mathbf{X}_n' \mathbf{X}_n)^{-1} \mathbf{X}_n' \mathbf{y}.$$

Fie $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}}$ o estimație a lui $\boldsymbol{\varepsilon}$. Atunci

$$\boldsymbol{\varepsilon} - \hat{\boldsymbol{\varepsilon}} = \bar{\varepsilon} \mathbf{1}_n + \mathbf{X}_n (\mathbf{X}_n' \mathbf{X}_n)^{-1} \mathbf{X}_n' \boldsymbol{\varepsilon} = \bar{\varepsilon} \mathbf{1}_n + \mathbf{Z}_n \mathbf{Z}_n' \boldsymbol{\varepsilon}, \tag{202}$$

unde

$$\mathbf{Z}_n = \mathbf{X}_n (\mathbf{X}_n' \mathbf{X}_n)^{-1/2}$$

Analogul teoremei 15 este:

TEOREMA 21 (Fazal (1977)). *Pentru cazul regresiei multiple considerat anterior, definim $g_n(u)$ și $\hat{g}_n(u)$ ca în (191) și (192) cu funcțiile W și W_n date de teorema 18.i).*

Presupunem că pentru $n \to \infty$, au loc următoarele condiții pentru $\mathbf{X}_n' = \{X_{kj}(n)\}$ și $\mathbf{Z}_n = \{Z_{jk}(n)\}$, $1 \leqslant j \leqslant n$ și $1 \leqslant k \leqslant r$ pentru orice k cu r fixat

i) $$\sum_j X_{kj}(n) = 0$$

și pentru unii $\delta > 0$

ii) $$\lim_n \frac{\sum_j |Z_{kj}(n)|^{2+\delta}}{(\sum_j Z_{kj}^2(n))^{1+\delta/2}} = 0$$

iii) $$M|\varepsilon_1|^{2+\delta} < \infty.$$

Atunci au loc concluziile a), b) și c) ale teoremei 18.

Scanned with OKEN Scanner

Demonstrație. Fie

$$R_n(u) = \hat{g}_n(u) - g_n(u). \tag{203}$$

Atunci ca în teorema 18 și folosind (202) obținem

$$|R_n(u) \leqslant C_\lambda\, n^{-1}\, b_n^{-1-\lambda} \sum_i |\varepsilon_i - \bar{\varepsilon}_i|^\lambda \leqslant$$

$$\leqslant C_\lambda\, n^{-\lambda/2}\, b^{-1-\lambda}\{|n^{1/2}\,\varepsilon| + n^{-1+\lambda/2} \sum_i |\sum_k Z_{ik}(n) \sum_j Z_{jk}(n)\varepsilon_j|^\lambda\} = A_n + B_n.$$

Indicii i, j merg de la 1 pînă la n, iar k de la 1 pînă la r. Urmează că:

$$A_n = \mathbf{0}_p(n^{-\lambda/2}\, b_n^{-1-\lambda}),$$

$$B_n \leqslant C_\lambda\, n^{-1}\, b_n^{-1-\lambda}\, r^{-1+\lambda} \sum_k (\sum_i |Z_{ik}(n)|^\lambda | \sum_j Z_{jk}(n)\varepsilon_j|^\lambda) \leqslant$$

(204)

$$\leqslant C_\lambda\, n^{-1}\, b_n^{-1-\lambda}\, r^{-1+\lambda}\{\sum_k (\sum_i |Z_{ik}(n)|^\lambda)^2\}^{1/2}\{\sum_k |\sum_j Z_{jk}(n)\varepsilon_j|^{2\lambda}\}^{1/2}.$$

Considerăm

$$\sum_k (\sum_i |Z_{ik}(n)|^\lambda)^2 \leqslant n^{2-\lambda} \sum_k (\sum_i Z_{ki}^2(n))^\lambda \leqslant n^{2-\lambda}\, r^{1-\lambda}\, (\sum_k \sum_i Z_{ik}^2(n))^\lambda =$$

$$= n^{2-\lambda}\, r^{1-\lambda}\, (\mathrm{Ur}\, \mathbf{Z}_n'\mathbf{Z}_n)^\lambda = n^{2-\lambda}\, r.$$

Astfel:

$$B_n \leqslant C_\lambda\, n^{-\lambda/2}\, b_n^{-1-\lambda}\, r^{-1/2+\lambda}\{\sum_k |\sum_j Z_{jk}(n)\varepsilon_j|^{2\lambda}\}^{1/2}.$$

Pentru fiecare $k = 1, \ldots, r$ condiția lui Liapunov pentru teorema limită centrală pentru variabilele independente $\{Z_{jk}(n)\varepsilon_j,\ j = 1, \ldots, n\}$ este satisfăcută de condițiile ii) și iii). Deci $B_n = \mathbf{0}_p(n^{-\lambda/2}\, b_n^{-1-\lambda})$. Primul rezultat este obținut din (203), (204) și rezultatele anterioare. Celelalte rezultate sînt demonstrate în același mod ca în teorema 17.

5.5. Convergența aproape sigură

Rezultate mai puternice decît în § 5.2 se pot obține dacă considerăm convergența aproape sigură a lui $\hat{g}_n(u)$.

Fazal (1977) folosește următorul rezultat care poate fi obținut după metoda lui Stout (1968).

TEOREMA 22. *Dacă pentru un* δ *pozitiv,* $\delta < \frac{1}{2}$, $M|\varepsilon_1|^{2/(1-2\delta)} < \infty$, *atunci*

$$n^{-\delta} \left| \sum_{j=1}^{n} z_{jn}\, \varepsilon_j \right| \xrightarrow{\text{a.s.}} 0$$

Scanned with OKEN Scanner

şi

$$n^{1/2}(\log\log n)^{-1/2}\,|\,\bar{\varepsilon}\,| \xrightarrow{\text{a.s.}} 0.$$

Un rezultat mai puternic decît cel din teorema 18 este

TEOREMA 23 (Fazal (1977)). *Fie $g_n(u)$ şi $\hat{g}_n(u)$ definite de* (191) *şi* (192). *Fie W funcţia de pondere cu variaţie mărginită şi care satisface uniform condiţia lui Lipschitz de ordinul $\lambda \in (0, 1)$. Dacă $M|\varepsilon_1|^{2/(1-2\delta)} < \infty$ pentru $0<\delta<1/2$, atunci*

$$\text{a)}\quad \sup_u |\hat{g}_n(u) - g_n(u)| = o_{\text{a.s.}}(n^{-\lambda/2}(\log\log n)^{\lambda/2} b_n^{-1-\lambda}) +$$

$$+ o_{\text{a.s.}}(n^{-\lambda/2+\lambda\delta} b_n^{-1-\lambda}).$$

Mai mult, dacă pentru fiecare real pozitiv γ seria

$$\sum_n \exp\{-\gamma n b_n^2/\alpha_n^2\}$$

converge cu $\alpha_n^{-1} = o(1)$, $\alpha_n b_n = o(1)$ şi $n^{-1} b_n^{-2} \alpha_n^2 = o(1)$, atunci

$$\text{b)}\quad \sup_u |\hat{g}_n(u) - g(u)| = o_{\text{a.s.}}(n^{-\lambda/2}(\log\log n)^{\lambda/2} b_n^{-1-\lambda}) +$$

$$+ o_{\text{a.s.}}(n^{-\lambda/2+\lambda\delta} b_n^{-1-\lambda}) + O_{\text{a.s.}}(\alpha_n^{-1}).$$

Demonstraţie. Cu aceleaşi notaţii ca în teorema 18, după inegalitatea (195) avem

$$\sup_u |R_n(u)| = \sup_u |\hat{g}_n(u) - g_n(u)| \leqslant A_n + B_n,$$

unde

$$A_n = C_\lambda(n^{-\lambda/2}(\log\log n)^{\lambda/2} b_n^{-1-\lambda}|\, n^{1/2}(\log\log n)^{-1/2}\bar{\varepsilon}|^\lambda.$$

şi

$$B_n \leqslant C_\lambda n^{-\lambda/2+\lambda\delta} b_n^{-1-\lambda}|\, n^{-\delta}\sum_j z_{jn}\varepsilon_j|^\lambda.$$

După teorema 22

$$A_n = o_{\text{a.s.}}(n^{-\lambda/2}(\log\log n)^{\lambda/2} b_n^{-1-\lambda})$$

şi

$$B_n = o_{\text{a.s.}}(n^{-\lambda/2+\lambda\delta} b_n^{-1-\lambda})$$

Cu acestea urmează partea a) a teoremei.
Folosind teorema 18 şi partea a) a teoremei prezente, urmează partea b. Deci teorema este demonstrată.

Observaţie. Din cauza condiţiei $\alpha_n b_n = o(1)$, nu se pot da rezultate similare pentru cazul rectangular al funcţiei de pondere.

Scanned with OKEN Scanner

Capitolul XXXVII

REGRESIE ROBUSTĂ. VERIFICĂRI DE IPOTEZE

§ 1. Teste de rang pentru parametrii mai multor drepte de regresie

1.1. Teste de rang bazate pe rangurile individuale a k selecții diferite pentru paralelismul dreptelor de regresie

1.1.1. Notații. Presupuneri

Pentru fiecare întreg pozitiv ν $(1 \leqslant \nu < \infty)$ considerăm șirul N_ν $\left(= \sum_{i=1}^{k} n_{\nu i}\right)$ variabile aleatoare independente $Y_{\nu ij}$, $j=1, \ldots, n_{\nu i}$, $i=1, \ldots, k$, unde presupunem că

$$P[Y_{\nu ij} \leqslant x] = F_{\nu ij}(x) = F(x - \alpha - \beta_i c_{\nu ij}),\ F \in \mathscr{F} \tag{1}$$

unde : $\mathbf{c}_{\nu i} = (c_{\nu i1}, \ldots, c_{\nu i n_{\nu i}})$, $i = 1, \ldots, k$ sînt vectori de constante cunoscute

$\alpha_1, \ldots, \alpha_k$ (intercepțiile) sînt parametrii nedoriți

$\beta_1, \ldots, \beta_k$ (pantele) sînt parametrii regresiei și

$\mathscr{F}$ este clasa tuturor funcțiilor de repartiție absolut continui ce se bucură de proprietatea :

$$\mathscr{F} = \{F : I(F) < \infty\}.$$

Ne propunem să verificăm ipoteza :

$$H_0 : \beta_1 = \ldots = \beta_k = \beta \quad \text{(necunoscut)} \tag{2}$$

față de alternativa conform căreia nu toți $\beta_1, \ldots, \beta_k$ sînt egali.

Definim :

$$\bar{c}_{\nu i} = n_{\nu i}^{-1} \sum_{j=1}^{n_{\nu i}} c_{\nu ij},\ C_{\nu i}^2 = \sum_{j=1}^{n_{\nu i}} (c_{\nu ij} - \bar{c}_{\nu i})^2,\ i = 1, \ldots, k \tag{3}$$

$$C_\nu^2 = \sum_{i=1}^{k} C_{\nu i}^2 \text{ și } \gamma_{\nu i} = C_{\nu i}^2 / C_\nu^2 \text{ pentru } i = 1, \ldots, k \tag{4}$$

Scanned with OKEN Scanner

Presupunem că pentru $\nu \to \infty$, $n_{\nu i}$, $C_{\nu i}^2$, $i = 1, \ldots, k$ tind la ∞ satisfăcînd

$$\gamma_{\nu i} \to \gamma_i : 0 < \gamma_0 \leqslant \gamma_1, \ldots, \gamma_k \leqslant 1 - \gamma_0 \leqslant 1, \text{ unde } \gamma_0 \leqslant 1/k \tag{5}$$

și condiția lui Noether

$$\lim_{\nu \to \infty} [\max_{1 \leqslant j \leqslant n_{\nu i}} |c_{\nu ij} - \bar{c}_{\nu i}|/C_{\nu i}] = 0 \text{ pentru toți } i = 1, \ldots, k. \tag{6}$$

Mai departe, fără a pierde din generalitate, presupunem că pentru $\nu (1 \leqslant \nu < \infty)$

$$c_{\nu i1} \leqslant c_{\nu i2} \leqslant \ldots \leqslant c_{\nu i n_{\nu i}}. \tag{7}$$

Fie $\varphi(u)$ o funcție absolut continuă și nedescrescătoare de $u : 0 < u < 1$ și presupunem că $\varphi(u)$ este de patrat integrabilă pe (0,1).

Definim

$$a_{\nu,j}^{(i)} = M[\varphi(U_{\nu j})], \quad j = 1, \ldots, n_{\nu i}, \; i = 1, \ldots, k. \tag{8}$$

Alternativ, putem defini scorurile

$$a_{\nu,j}^{(i)} = \varphi(j/[n_{\nu i} + 1]), \quad j = 1, \ldots, n_{\nu i}, \quad i = 1, \ldots, k. \tag{9}$$

Două scoruri particulare sînt folosite

- $\varphi(u) = u : 0 < u < 1$ și atunci scorurile din (8) sau (9) sînt date de $j/[n_{\nu i} + 1]$, $j = 1, \ldots, n_{\nu i}$ și sînt cunoscute ca scoruri Wilcoxon
- $\varphi(u)$ inversa funcției lui Laplace în u și sînt cunoscute sub numele de scoruri normale.

Fie

$$\bar{a}_{\nu}^{(i)} = (1/n_{\nu i}) \sum_{j=1}^{n_{\nu i}} a_{\nu,j}^{(i)}, \quad i = 1, \ldots, k \tag{10}$$

și A^2 și $\bar{\varphi}$ definite ca în (XXXVI.121).
Considerăm statistica

$$T_{\nu,i} = \left[\sum_{j=1}^{n_{\nu i}} (c_{\nu ij} - \bar{c}_{\nu i}) a_{\nu, R_{\nu ij}}^{(i)}\right] \Big/ \left[A \cdot C_{\nu i}\right], \quad i = 1, \ldots, k, \tag{11}$$

unde $R_{\nu ij}$ este rangul lui $Y_{\nu ij}$ pentru $Y_{\nu i1}, \ldots, Y_{\nu i n_{\nu i}}$ și $C_{\nu i}$ este definit de (3).

Observație. Astfel de statistici sînt studiate de Hoeffding (1950), Terry (1952) și Hájek (1962) pentru verificarea ipotezei privind regresia în cazul unei singure selecții.

Fie

$$T_{\nu,i} = T_{\nu,i}(\mathbf{Y}_{\nu i}), \text{ unde } \mathbf{Y}_{\nu i} = (Y_{\nu i1}, \ldots, Y_{\nu i n_{\nu i}}), \quad i = 1, 2, \ldots, k. \tag{12}$$

$T_{\nu,i}(\mathbf{Y}_{\nu i} + bc_{\nu i})$ statistica (11) cînd observațiile $Y_{\nu ij}$ sînt înlocuite prin $Y_{\nu ij} + bc_{\nu ij}$, unde b este un număr real.

Scanned with OKEN Scanner

Fie $\varphi(u, f)$ definită ca în (XXXVI. 119) și $I(F) < \infty$
Fie

$$\rho(\varphi(u,f),\ \varphi) = \left[\int_0^1 \varphi(u,f)\varphi(u)du\right]/[A^2 I(F)]^{1/2} \tag{13}$$

și presupunem că $\rho(\varphi(u, f), \varphi)$ este strict pozitiv. Această proprietate de monotonicitate a lui $\varphi(u)$ are loc dacă luăm

$$\varphi(u) = -[g'(G^{-1}(u))/g(G^{-1}(u))],\ 0 < u < 1,\ G \in \mathcal{F} \tag{14}$$

unde $g(x)$ este o densitate puternic unimodală. Dacă $f(x)$ și $g(x)$ sînt puternic unimodale, $\rho(\varphi(u, f), \varphi)$ oricare ar fi A^2 și $I(F)$.

În cele ce urmează, va fi necesar să considerăm un estimator consistent pentru $\int_0^1 \varphi(u, f)\cdot\varphi(u)du$ și pentru aceasta vom cere ca

(15) $(\mathrm{d}/\mathrm{d}x)\ \varphi(F(x))$ este continuă și mărginită pentru toți $x(-\infty < x < \infty)$

1.1.2. Teste parametrice asimptotic optime

Sub (1) fie $P(\mathbf{Y}_\nu;\ \boldsymbol{\alpha},\ \boldsymbol{\beta})$ funcția de verosimilitate a variabilelor $\mathbf{Y}_{\nu 1}, \ldots, \mathbf{Y}_{\nu k}$, unde $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ și $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$. Fie Ω_{2k} spațiul real $2k$-dimensional al parametrilor $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ și fie $\Omega^*_{(k+1)}$ subspațiul lui Ω_{2k} pentru care $\beta_1 = \ldots = \beta_k$.
Statistica testului raportului de verosimilități este dată de

$$l_\nu = [\sup_{\Omega^*_{(k+1)}} P(\mathbf{Y}_\nu;\ \boldsymbol{\alpha},\ \boldsymbol{\beta}) / \sup_{\Omega_{2k}} P(\mathbf{Y}_\nu,\ \boldsymbol{\alpha},\ \boldsymbol{\beta})].$$

Din teorema XIV.3, $-2\log l_\nu$ are asimptotic o repartiție $\chi^2(k-1)$.
Deci, pentru selecții de volum mare, testul raportului de verosimilități constă în următoarele:
respingem ipoteza (2) dacă

$$-2\log l_\nu > \chi^2_\varepsilon\ (k-1), \tag{16}$$

unde $\chi^2_\varepsilon(k-1)$ este ε-punctul superior al repartiției $\chi^2(k-1)$

Fie $\hat{\beta}_{\nu i}$ estimatorul de verosimilitate maximă pentru β_i, $i = 1, \ldots, k$ și fie

$$Q = I(F)\left[\sum_{i=1}^{k} C^2_{\nu i}(\hat{\beta}_{\nu i} - \hat{\beta}_{\nu 0})^2\right],\ \hat{\beta}_{\nu 0} = \sum_{i=1}^{k} \gamma_{\nu i}\,\beta_{\nu i}. \tag{17}$$

Urmează că

$$Q_\nu + 2\log l_\nu \xrightarrow[\nu\to\infty]{P} 0. \tag{18}$$

Scanned with OKEN Scanner

Deoarece $\hat{\beta}_{\nu i}$ sînt estimaţii consistente pentru β_i şi $C_{\nu i}^2 \xrightarrow[\nu\to\infty]{} \infty$, pentru toţi $i = 1, \ldots, k$, Q_ν poate fi făcut arbitrar de mare (cînd $\nu \to \infty$) cînd nu toţi $\beta_1, \ldots, \beta_k$ sînt egali. Deci testul din (16) este consistent. Astfel pentru a studia proprietăţile sale asimptotice vom considera următorul şir de ipoteze alternative pentru care puterea testului este întotdeauna mărginită de 1.

$$H_\nu: \ \beta_i = \beta + C_\nu^{-1}\theta_i, \quad i = 1, \ldots, k; \quad \sum_{i=1}^{k} \gamma_i \theta_i = 0, \tag{19}$$

unde $\theta_1, \ldots, \theta_k$ sînt toţi reali şi β este un parametru nedorit. Atunci sub H_ν, $-2\log l_\nu$ (sau Q_ν) are asimptotic o repartiţie χ^2 necentrată cu $k-1$ grade de libertate şi parametru de necentricitate

$$\Delta_Q = I(F)\left[\sum_{i=1}^{k} \gamma_i \theta_i^2\right], \tag{20}$$

unde $\gamma_1, \ldots, \gamma_k$ sînt definiţi de (5).

Acum, pentru fiecare a pozitiv şi fiecare punct $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in R^k$ definim suprafaţa $S_a(\boldsymbol{\theta})$ prin ecuaţia

$$S_a(\boldsymbol{\theta}) = \left\{\boldsymbol{\theta} : \sum_{i=1}^{k} \gamma_i \theta_i^2 = a\right\}. \tag{21}$$

Considerăm transformarea $\boldsymbol{\theta}^* = \mathbf{D}\boldsymbol{\theta}$, unde prima linie a lui $\mathbf{D}$ este $(\gamma_1, \ldots, \gamma_k)$ şi

$$\mathbf{D}\mathbf{D}' = \boldsymbol{\Gamma} = \operatorname{diag}(\gamma_1, \ldots, \gamma_k). \tag{22}$$

Această transformare duce suprafaţa $S_a(\boldsymbol{\theta})$ în sfera dată de

$$\sum_{i=2}^{k} (\theta_i^*)^2 = a, \text{ unde prin definiţie } \theta_1^* = 0. \tag{23}$$

Pentru orice punct $\boldsymbol{\theta}_0$ şi orice δ pozitiv, considerăm mulţimea $\omega(\boldsymbol{\theta}_0, \delta)$ ce constă în toate punctele $\boldsymbol{\theta}$ care variază pe aceeaşi $S_a(\boldsymbol{\theta})$ ca $\boldsymbol{\theta}_0$ şi pentru care $|\boldsymbol{\theta} - \boldsymbol{\theta}_0| < \delta$. Fie

$$\eta(\boldsymbol{\theta}) = \lim_{\delta\to 0} \{\mathscr{A}[\omega'(\boldsymbol{\theta}, \delta)]/\mathscr{A}[\omega(\boldsymbol{\theta}, \delta)]\}, \tag{24}$$

unde $\omega'(\boldsymbol{\theta}, \delta)$ este imaginea lui $\omega(\boldsymbol{\theta}, \delta)$ prin transformarea $\boldsymbol{\theta}^* = \mathbf{D}\boldsymbol{\theta}$ şi $\mathscr{A}(\omega)$ reprezintă aria mulţimii ω. Atunci are loc.

TEOREMA 1 (Sen (1969). *Testul raportului de verosimilităţi* (16) *pentru verificarea ipotezei* (2) *se bucură de proprietăţile*

i) *are asimptotic cea mai bună putere medie în raport cu suprafeţele* $S_a(\boldsymbol{\theta})$ *şi funcţiile de pondere* $\eta(\boldsymbol{\theta})$.

Scanned with OKEN Scanner

ii) *are asimptotic cea mai bună putere constantă pe suprafețele* $S_a(\theta)$ *și*

iii) *este asimptotic cel mai stringent test.*

În baza teoremei 1 putem privi testul raportului de verosimilități din (16) ca testul parametric optim asimptotic. În practică cel mai folosit test este testul raportului dispersiilor conținînd estimatorii de C.M.M.P.

$$Z_\nu = \left[\sum_{i=1}^{k} C_{\nu i}^2(\bar{\beta}_{\nu i} - \bar{\beta}_{\nu 0})^2/(k-1)s_e^2\right] \tag{25}$$

unde $\bar{\beta}_{\nu i} = \sum_{j=1}^{n_{\nu i}} Y_{\nu ij}(c_{\nu ij} - \bar{c}_{\nu i})/C_{\nu i}^2,\ i = 1, \ldots, k.$

$$\bar{\beta}_{\nu 0} = \sum_{i=1}^{k} \gamma_{\nu i}\bar{\beta}_{\nu i}.$$

Pentru orice $F(x)$ avînd dispresia finită $\sigma^2(F)$, asimptotic sub (2) [(19)].

$$(k-1)Z_\nu \sim \chi^2(k-1)[\sim \chi^2(k-1, \Delta_z)],$$

unde

$$\Delta_z = \left[\sum_{i=1}^{k} \gamma_i\theta_i^2\right]/\sigma^2(F). \tag{26}$$

Sub (19), $-2\log l_\nu$ și $(k-1)Z_\nu$ au asimptotic repartiții χ^2 cu $k-1$ grade de libertate și parametrii de necentricitate Δ_Q și respectiv Δ_z. Deci eficiența asimptotică relativă a testului Z_ν în raport cu testul l_ν este

$$e_{z,l} = \Delta_z/\Delta_Q = [\sigma^2(F)I(F)]^{-1} \leqslant 1,$$

după inegalitatea lui Rao-Cramer.

1.1.3. Teste de rang asimptotic cel mai puternice (C.M.P.)

Din (17) și (25) observăm că l_ν și Z_ν sînt bazate pe estimatori de verosimilitate maximă și respectiv de C.M.M.P. pentru $\beta_1, \ldots, \beta_k$. Sen (1969) consideră o estimație ponderată a coeficientului de regresie. În acest scop definește

$$T_\nu^* = \sum_{i=1}^{k} C_{\nu i}T_{\nu i}/C_\nu = \left[\sum_{i=1}^{k}\sum_{j=1}^{n_{\nu i}} (c_{\nu ij} - \bar{c}_{\nu i})a_{\nu,k_{\nu ij}}^{(i)}\right]/[AC_\nu]. \tag{27}$$

Vom exprima T_ν^* ca $T_\nu^*(\mathbf{Y}_\nu)$ și scriem $\mathbf{c}_\nu = (c_{\nu 1}, \ldots, c_{\nu k})$. Avem :

$$T_\nu^*(\mathbf{Y}_\nu - b\mathbf{c}_\nu) \text{ este } \downarrow \text{ în } b\ (-\infty < b < \infty). \tag{28}$$

De asemenea din (8), (9) şi (11) urmează că

$$M(T_{\nu,i}(\mathbf{Y}_{\nu,i}) \mid \beta_i = 0) = 0, \text{ pentru toţi } i = 1, \ldots, k. \tag{29}$$

În consecinţă, sub (2)

$$M(T_\nu^*(\mathbf{Y}_\nu - \beta \mathbf{c}_\nu) \mid \beta_1 = \ldots = \beta_k = \beta) = 0. \tag{30}$$

Notăm că funcţia de repartiţie a lui $T_\nu^*(\mathbf{Y}_\nu - \beta \mathbf{c}_\nu)$ poate sau nu să fie simetrică faţă de zero. O condiţie suficientă pentru această simetrie este ca

$$\varphi(u) + \varphi(1 - u) = 2\,\varphi(1/2)$$

pentru toţi $0 < u < 1$. Însă în cele ce urmează nu impunem simetrie nici pentru $\varphi(u)$ nici pentru $\varphi(u, f)$.

Vom estima β (sub (2)) din

$$T_\nu^*(\mathbf{Y}_\nu - b\mathbf{c}_\nu) = 0.$$

Pentru aceasta fie

$$\begin{aligned} \beta_{\nu(1)}^* &= \sup \{b : T_\nu^*(\mathbf{Y}_\nu - b\mathbf{c}_\nu) > 0\} \\ \beta_{\nu(2)}^* &= \inf\{b : T_\nu^*(\mathbf{Y}_\nu - b\mathbf{c}_\nu) < 0\}. \end{aligned} \tag{31}$$

Atunci estimatorul ponderat pentru β este

$$\beta_\nu^* = (\beta_{\nu(1)}^* + \beta_{\nu(2)}^*)/2. \tag{32}$$

Definim

$$\hat{T}_{\nu,i} = T_{\nu,i}(\mathbf{Y}_{\nu,i} - \beta_\nu^* \cdot \tilde{c_{\nu i}}), \quad i = 1, 2, \ldots, k \text{ şi } L_\nu = \sum_{i=1}^{k} \hat{T}_{\nu,i}^2. \tag{33}$$

Are loc -

TEOREMA 2 (Sen (1969)). *Sub* (2) *şi condiţiile din* § 1.1.1., *asimptotic*

$$L_\nu \sim \chi^2(k - 1).$$

În baza teoremei 2 avem următorul test asimptotic de volum ε:

(34) *respingem* H_0(*din* (2)) *dacă* $L_\nu > \chi_\varepsilon^2(k-1)$ *unde* $\chi_\varepsilon^2(k-1)$ *este definit de* (16).

TEOREMA 3. (Sen (1969)). *În condiţiile* § 1.1.1 *şi şirul de alternative din* (19), L_ν *are asimptotic o repartiţie* χ^2 *cu* $k - 1$ *grade de libertate şi parametru de necentricitate*

$$\Delta_L = [\rho(\varphi(u, f), \overset{*}{\varphi})]^2 I(F) \left[\sum_{i=1}^{k} \gamma_i \theta_i^2\right] \tag{35}$$

unde $\rho(\varphi(u, f), \varphi)$ *este definit de* (13).

Scanned with OKEN Scanner

1.1.4. Test pentru un model mai general

În locul lui (1) considerăm următorul model

$$(36)\quad P[\mathbf{Y}_{\nu i j} \leqslant x] = F_i(x - \alpha_i - \beta_i c_{\nu i j}),\ j = 1, \ldots, n_{\nu i},\ i = 1, \ldots, k$$

pentru toți $F_i \in \mathscr{F}$ pentru toți $i = 1, \ldots, k$ și $c_{\nu i j}$ satisfăcînd condițiile § 1.1.1. Aici verificăm ipoteza nulă (2) fără a presupune că $F_1, \ldots, F_k$ au aceeași formă.

Definim statisticile $T_{\nu,i}$ ca în (11) și de asemenea estimatorii individuali $\beta^*_{\nu,i}$ ca în § 1.1.3.

Acum în ipoteza conform căreia $\beta_i = 0$ este posibil să localizăm două valori $T^{(1)}_{\nu,i}$ și $T^{(2)}_{\nu,i}$ astfel încît:

$$(37)\quad P[T^{(1)}_{\nu,i} \leqslant T_{\nu,i} \leqslant T^{(2)}_{\nu,i} | \beta_i = 0] = 1 - \alpha, \text{ pentru toți } F_i \in \mathscr{F},$$

unde $\alpha (0 < \alpha < 1)$ este prestabilit. Atunci definim

$$(38)\quad \beta^*_{\nu,i,L} = \inf \{b : T_{\nu,i}(\mathbf{Y}_{\nu i} - b c_{\nu i}) \leqslant T^{(2)}_{\nu,i}\}$$

$$(39)\quad \beta^*_{\nu,i,U} = \sup \{b : T_{\nu,i}(\mathbf{Y}_{\nu i} - b c_{\nu i}) \geqslant T^{(1)}_{\nu,i}\}.$$

Astfel din (28), (36), (37), (38) și (39) obținem

$$(40)\quad P[\beta^*_{\nu,i,L} \leqslant \beta_i \leqslant \beta^*_{\nu,i,U} | \beta_i] = 1 - \alpha.$$

Vom folosi aceste intervale de încredere pentru a estima anumiți parametrii. În (XXXVI.119), dacă înlocuim F prin F_i, scorul corespunzător va fi notat prin $\varphi_i(u, f)$ și în (13) înlocuind $\varphi(u, f)$ prin $\varphi(u, f)$, definim $\rho(\varphi_i(u, f), \varphi) i = 1, \ldots, k$. Trebuie să estimăm parametrii

$$(41)\quad B_i = B(\varphi_i(u, f), \varphi) = \left[\int_0^1 \varphi_i(u, f)\varphi(u)du\right],\quad i = 1, \ldots, k.$$

Pentru aceasta, ținînd seama că sub $\beta_i = 0$, $T_{\nu,i}$ este asimptotic normală, avem (folosind cozi simetrice în (37))

$$(42)\quad T^{(j)}_{\nu,i} \underset{k\to\infty}{\to} (-1)^j z_{\alpha/2},\ \text{pentru } j = 1, 2 \text{ și } i = 1, \ldots, k$$

unde $z_{\alpha/2}$ este α punctul superior al repartiției $N(0, 1)$.

TEOREMA 4 (Sen (1969)). *Pentru modelul* (36) *și în condițiile* § 1.1.1

$$(43)\quad B^{\wedge}_{\nu i} = A[T^{(2)}_{\nu,i} - T^{(1)}_{\nu,i}]/[C_{\nu i}(\beta^*_{\nu,i,U} - \beta^*_{\nu,i,L})] \underset{P}{\to} B_i,\quad i = 1, \ldots, k.$$

TEOREMA 5 (Sen (1969)). *Sub* (36) *și condițiile* § 1.1.1., *variabilele aleatoare independente*

$$B_i C_{\nu i}(\beta^*_{\nu,i} - \beta_i)/A,\ i = 1, \ldots, k$$

sînt asimptotic normale cu mediile zero și dispersia unitate.

Scanned with OKEN Scanner

Acum definim

$$\gamma_{\nu i}^{*} = C_{\nu,i}^{2}\hat{B}_{\nu,i}^{2}/C_{\nu}^{2},\ i = 1, \ldots, k;\ \beta_{\nu}^{**} = \sum_{i=1}^{k} \gamma_{\nu i}^{*}\beta_{\nu,i}^{*}/\sum_{i=1}^{k}\gamma_{\nu i}^{2}. \tag{44}$$

Atunci, testul propus este bazat pe statistica

$$S_{\nu} = [\sum_{i=1}^{k} C_{\nu i}^{2}\hat{B}_{\nu i}^{2}(\beta_{\nu,i}^{*} - \beta_{\nu}^{**})^{2}]/A^{2}. \tag{45}$$

TEOREMA 6 (Sen (1969)). *Sub H_0 în (2) și în condițiile* § 1.1.1., S_ν *are asimptotic o repartiție* $\chi^2(k-1)$.

În virtutea acestui rezultat, poate fi construită ca în (34) o repartiție asimptotică liberă pentru H_0 din (2) corespunzătoare modelului (36). Acum, după (4), (5) și (43) urmează că

$$\gamma_{\nu i}^{*} \xrightarrow[\nu\to\infty]{P} \gamma_{i}^{*},$$

unde

$$\gamma_{i}^{*} = \gamma_{i}B_{i}^{2},\ i = 1, \ldots, k. \tag{46}$$

Definim $\theta_1, \ldots, \theta_k$ ca în (19) și fie $\bar{\theta} = \sum_{i=1}^{k} \gamma_i^{*}\theta_i/\sum_{i=1}^{k}\gamma_i^{*}$. Atunci din teorema 5, (43), (46) și condițiile § 1.1.1., urmează că S_ν are asimptotic o repartiție χ^2 cu $k-1$ grade de libertate și parametru de necentricitate

$$\Delta_s' = \left[\sum_{i=1}^{k} \gamma_i^{*}(\theta_i - \bar{\theta})^{2}\right]/A^{2}. \tag{47}$$

(dacă $\{H_{\nu i}\}$ din (19) sînt adevărate).

Dacă funcțiile de repartiție $F_1, \ldots, F_k$ în (36) sînt toate cunoscute, putem construi testul raportului de verosimilitate ca în § 1.1.2. În acest caz, statistica testului raportului de verosimilitate va fi notată prin l_ν^*. Atunci testul poate fi construit ca în (16), și de asemenea, sub (19), $-2\log l_\nu^*$ are asimptotic o repartiție χ^2 necentrată cu $k-1$ grade de libertate și parametru de necentricitate

$$\Delta_0^{*} = \left[\sum_{i=1}^{k} \gamma_i(\theta_i - \bar{\theta}^{*})^{2} I(F_i)\right], \tag{48}$$

unde

$$\bar{\theta}^{*} = \sum_{i=1}^{k} I(F_i)\gamma_i\theta_i/\sum_{i=1}^{k} I(F_i)\gamma_i.$$

Prin urmare, notînd că

$$\sum_{i=1}^{k} \gamma_i^{*}(\theta_i - \bar{\theta})^{2} \leqslant \sum_{i=1}^{k}\gamma_i(\theta_i - \bar{\theta}^{*})^{2}; \tag{49}$$

$$\gamma_i^{*} = \gamma_i B_i^{2} = \gamma_i I(F_i)A^{2}\rho^{2}(\varphi_i(u, f), \varphi) \text{ pentru toți } i = 1, \ldots, k, \tag{50}$$

Scanned with OKEN Scanner

din (47) împreună cu (50), obţinem

$$(51)\qquad \Delta_s/\Delta_Q^* \leqslant \frac{\sum_{i=1}^{k} \gamma_i I(F_i)\,\rho^2(\varphi_i(u,f),\ \varphi)(\theta_i - \bar{\theta}^*)^2}{\sum_{i=1}^{k} \gamma_i I(F_i)(\theta_i - \bar{\theta}^*)^2} \leqslant 1$$

şi egalitatea are loc numai cînd $\rho(\varphi_i(u,f),\ \varphi) = 1$ pentru toţi $i = 1, \ldots, k$.

Cazul particular cînd (51) este independentă de $\theta_1, \ldots, \theta_k$ este de o importanţă deosebită. Considerăm modelul heteroscedastic:

$$(52)\qquad F_i(x) = F(x/\sigma_i) \quad \text{pentru toţi} \quad i = 1, \ldots, k$$

unde $\sigma_1, \ldots, \sigma_k$ sînt factori de scară pozitivi. Atunci

$$(53)\qquad I(F_i) = I(F)/\sigma_i^2 \text{ şi } \rho(\varphi_i(u,f),\ \varphi) = \rho\,(\varphi(u,f),\ \varphi) \text{ pentru toţi } i = 1, \ldots, k.$$

unde $\rho(\varphi(u,f),\ \rho)$ este definită în (13). În acest caz urmează că (51) este egal cu $\rho^2(\varphi(u,f),\ \varphi)$ independent de $\theta_1, \ldots, \theta_k$ şi $\sigma_1, \ldots, \sigma_k$.

Acum considerăm estimatorii de C.M.M.P. $\hat{\beta}_{vi}$, $i = 1, \ldots, k$ definiţi ca în (25). De asemenea fie $s_{e,i}^2$ dispersia datorită erorii de la selecţia i, $i = 1, \ldots, k$ şi

$$(54)\qquad \hat{\gamma}_{vi}^0 = \gamma_{vi}/s_{e,i}^2, \quad i = 1, \ldots, k\,.$$

Astfel după (4), (5) şi proprietatea de convergenţă a dispersiei $s_{e,i}^2$, urmează că

$$(55)\qquad \hat{\gamma}_{vi}^0 \xrightarrow[v\to\infty]{P} \gamma_i/\sigma_i^2 = \gamma_i^0 \text{ pentru toţi } i = 1, \ldots, k.$$

Fie

$$\bar{\beta}_v^* = \sum_{i=1}^{k} \gamma_{vi}^0 \bar{\beta}_{vi} \Big/ \sum_{i=1}^{k} \gamma_{vi}^0 \text{ şi } \bar{\theta}^{**} = \sum_{i=1}^{k} \gamma_i^0 \theta_i \Big/ \sum_{i=1}^{k} \gamma_i^0.$$

Urmează că

$$(56)\qquad Z_v^* = C_v^2\left[\sum_{i=0}^{k} \gamma_{vi}^0(\bar{\beta}_{vi} - \bar{\beta}_v^*)^2\right],$$

are asimptotic [sub (2) şi (36)] o repartiţie $\chi^2(k-1)$ şi deci testul pentru selecţii de volum mare poate fi construit ca în (16).

Sub (19), Z_v^* urmează asimptotic o repartiţie χ^2 cu $k-1$ grade de libertate şi parametru de necentricitate

$$(57)\qquad \Delta_z^* = \left[\sum_{i=1}^{k} \gamma_i^0(\theta_i - \bar{\theta}^{**})^2\right].$$

Scanned with OKEN Scanner

Notînd că

$$\sum_{i=1}^{k} \gamma_i^0(\theta_i - \bar{\theta}^{**}) \leqslant \sum_{i=1}^{k} \gamma_i^0(\theta_i - \bar{\theta})^2 \tag{58}$$

și că

$$\gamma_i^* = \gamma_i B_i^2 = \gamma_i^0 \sigma_i^2 B_i^2 \quad \text{pentru } i = 1, \ldots, k. \tag{59}$$

din (47), (57) împreună cu (59) obținem că eficiența asimptotică relativă (E.A.R.) a testului S_ν în raport cu testul Z_ν^* este egală cu

$$\Delta_s/\Delta_z^* \geqslant \frac{\sum_{i=1}^{k} \gamma_i^0 e_i(\theta_i - \bar{\theta})^2}{\sum_{i=1}^{k} \gamma_i^0(\theta_i - \bar{\theta})^2} \geqslant \min_i e_i, \tag{60}$$

unde

$$e_i = \sigma_i^2 B_i^2/A^2 = \text{EAR } (\beta_{\nu i}^*/\bar{\beta}_{\nu i}) \text{ pentru } i = 1, \ldots, k. \tag{61}$$

1.1.5. Monotonia statisticilor de rang de regresie

Fie $a_1 \leqslant \ldots \leqslant a_n$, n scoruri de rang și $c_1 \leqslant \ldots \leqslant c_n$ (nu toate egale) constante cunoscute. De asemenea fie $Y_1, \ldots, Y_n$ variabile aleatoare independente cu funcțiile de repartiție $F_1, \ldots,$ respectiv F_n continui. Fie $R_i(b)$ rangul lui $Z_i(b) = Y_i - bc_i$ printre $Z_1(b), \ldots, Z_n(b)$ pentru $i = 1, \ldots, n$.

Fie

$$T_n(b) = \sum_{i=1}^{n}(c_i - \bar{c})a_{R_i(b)}; \ \bar{c} = n^{-1}\sum_{i=1}^{n} c_i, \ -\infty < b < \infty. \tag{62}$$

TEOREMA 7 (Sen(1969)). *În condițiile anterioare, există* $n^*(1 \leqslant n \leqslant c_n^2)$ *pentru* $b_1 < \ldots < b_{n^*}$ *astfel încît*

i) $T_n(b) = T_n(b_s + 0)$ *pentru toți* $b_s < b < b_{s+1}$, $s = 0, 1, \ldots, n^*$ *unde* $b_0 = -\infty$, $b_{n^*+1} = \infty$

ii) $T_n(b_s - 0) \geqslant T_n(b_s) \geqslant T_n(b_s + 0)$, $s = 1, \ldots, n^*$ *și*

iii) $T_n(b)$ *este necesar pozitiv (negativ) pentru* $b < b_1 (> b_{n^*})$ *și astfel* $T_n(b)$ *este* $\downarrow$ *în* $b (-\infty < b < \infty)$.

1.2. Teste de rang pentru identitatea intercepțiilor mai multor drepte de regresie

1.2.1. Notații. Presupuneri

Considerăm o mulțime de $N \left(= \sum_{i=1}^{k} n_i \right)$ variabile aleatoare independente Y_{ij}, $1 \leqslant j \leqslant n_i$, $1 \leqslant i \leqslant k$, unde

$$P\{Y_{ij} \leqslant x\} = F_{ij}(x) = F(x - \alpha_i - \beta_i c_{ij}), \tag{63}$$

Scanned with OKEN Scanner

unde $\mathbf{c}_i = (c_{i1}, \ldots, c_{in_i})$, $i = 1, \ldots, k$ sînt vectori cunoscuți, iar β_i și α_i, $1 \leqslant i \leqslant k$ au semnificațiile din §1.

Ne propunem să verificăm ipoteza

$$H_0 : \alpha_1 = \ldots = \alpha_k = \alpha_0 \quad \text{(necunoscut)}, \tag{64}$$

față de mulțimea alternativelor că nu toți $\alpha_1, \ldots, \alpha_k$ sînt egali : deci în acest caz privim $\beta_1, \ldots, \beta_k$ ca parametrii nedoriți (nu necesar toți egali).

Vom presupune că $F \in \mathscr{F} = \{F : f(x) = f(-x), (\forall) x \geqslant 0$ și $I(F) < \infty\}$

Fie

$$\lambda_N^{(i)} = n_i/N, \quad 1 \leqslant i \leqslant k, \text{ (astfel încît } \sum_{i=1}^{k} \lambda_N^{(i)} = 1), \tag{65}$$

și presupunem că pentru $N \to \infty$

$$\lambda_N^{(i)} \to \lambda^{(i)}, \quad 0 < \lambda^{(i)} < 1, \quad i = 1, \; 2, \; \ldots, \; k. \tag{66}$$

Fără a pierde din generalitate, presupunem că pentru fiecare i

$$c_{i1} \leqslant \ldots \leqslant c_{in_i} \text{ cu cel puțin o inegalitate strictă, } 1 \leqslant i \leqslant k. \tag{67}$$

Fie

$$\bar{c}_{i,N} = (1/n_i) \sum_{j=1}^{n_i} c_{ij}, \; C_{i,N}^2 = \sum_{j=1}^{n_i} (c_{ij} - \bar{c}_{i,N})^2, \; q_{i,N}^2 = n_i \bar{c}_{i,N}^2 / C_{i,N}^2, \tag{68}$$

și presupunem că

$$\lim_{N \to \infty} C_{i,N}^2 = \infty \text{ pentru toți } 1 \leqslant i \leqslant k.$$

De asemenea presupunem că

$$\liminf n_i^{-1} C_{i,N}^2 \geqslant C_0^2 > 0, \tag{69}$$

$$\lim_{N \to \infty} [\max_{1 \leqslant j \leqslant n_i} |c_{ij} - \bar{c}_{i,N}| / C_{i,N}] = 0, \tag{70}$$

$$\limsup q_{i,N}^2 \leqslant q_0^2 < \infty \text{ pentru toți } i = 1, \ldots, k. \tag{71}$$

Nu presupunem că $n_i^{-1} C_{i,N}^2 \to C_i^2 > 0$ cînd $N \to \infty$.

Fie $\varphi(u)$, $0 < u < 1$, o funcție absolut continuă, nedescrescătoare și de patrat integrabilă pe $[0, 1]$. Presupunem că :

$$\varphi(u) + \varphi(1 - u) = 0, \quad \text{pentru toți } u : 0 < u < 1. \tag{72}$$

Definim o mulțime de n scoruri $a_n(j)$ ca în (XXXVI.117) și (XXXVI.118).

Scanned with OKEN Scanner

Fie A^2 și $\bar{\varphi}$ avînd semnificațiile din (XXXVI. 121) și

$$\bar{a}_n = \frac{1}{n}\sum_{j=1}^{n} a_n(j),\ A_n^2 = \frac{1}{n}\sum_{j=1}^{n}[a_n(j)-\bar{a}_n]^2. \tag{73}$$

Pentru estimarea parametrilor $\beta_1, \ldots, \beta_k$ necesară pentru alinierea observațiilor folosim următoarele statistici. Fie R_{ij}^0 rangul lui Y_{ij} printre $Y_{i1}, \ldots, Y_{in_i}$ și definim :

$$T_{i,N} = \left[\sum_{j=1}^{n_i}(c_{ij} - \bar{c}_{i,N})a_{n_i}(R_{ij}^0)\right][A_{n_i}^{-1}C_{i,N}^{-1}],\quad 1 \leqslant i < k. \tag{74}$$

De asemenea fie $T_{i,N}(b)$ statistica rang din (74) bazat pe observațiile $Y_{ij} - bc_{ij}$, $1 \leqslant j \leqslant n_i$. Din teorema 7 urmează că

(75) $T_{i,N}(b)$ este $\downarrow$ în b, $-\infty < b < \infty$ pentru toți $1 \leqslant i \leqslant k$.

Definim

$$\varphi^*(u) = \varphi((1+u)/2),\ 0 < u < 1,\ (A^*)^2 = \int_0^1 [\varphi^*(u)]^2\,du = A^2, \tag{76}$$

și considerăm mulțimea scorurilor

$$a_n^*(j) = M\{\varphi^*(u_{nj})\} \text{ sau } \varphi^*(j/(n+1)),\quad 1 \leqslant j \leqslant n. \tag{77}$$

Pentru a construi testul, considerăm următorul tip de statistică,

$$S_{i,N} = (1/n_i)\sum_{j=1}^{n_i} a_{n_i}^*(R_{ij}^+)\mathrm{Sgn}(Y_{ij})\quad 1 \leqslant i \leqslant k \tag{78}$$

unde R_{ij}^+ este rangul lui $|Y_{ij}|$ printre $|Y_{i1}|, \ldots, |Y_{in_i}|$. De asemenea dacă înlocuim în (78) Y_{ij} prin $Y_{ij} - a_i - b_ic_{ij}$, statisticile corespunzătoare sînt notate prin

$$S_{iN}(a_i, b_i),\quad 1 \leqslant i \leqslant k. \tag{79}$$

Statistica test este o formă patratică în $S_{i,N}(a_i,\ b_i)$, $1 \leqslant i \leqslant k$, unde b_i sînt aleși ca estimatori ai parametrilor β_i obținuți din $T_{i,N}(b_i)$ în (74) și $a_1 = \ldots = a_k = a$ este ales ca un estimator ponderat pentru α_0, valoarea ipotetică comună pentru α_i. Pentru b_i dat

(80) $S_{i,N}(a_i,\ b_i)$ este $\downarrow$ în a_i, $-\infty < a_i < \infty$, $1 \leqslant i \leqslant k$.

Cu $\varphi^+(u)$ definit ca în (76) și $\varphi^+(u, f)$ ca în (XXXVI.138) presupunem că

$$B(F,\ \varphi) = \int_0^1 \varphi(u)\ \varphi(u, f)\,du = \int_0^1 \varphi^*(u)\ \varphi^+(u,\ f)\,du > 0. \tag{81}$$

Dacă F este puternic unimodală și $\varphi^*(u)$ nu este constantă, atunci are loc (81).

1.2.2. Teste parametrice optime

În presupunerea de normalitate, cel mai folosit test este testul raportului dispersiilor:

$$Z_N = \left[\sum_{i=1}^{k} n_i(\tilde{\alpha}_i - \tilde{\tilde{\alpha}})^2/(1 + q_{i,N}^2)\right]/[(k-1)s_e^2], \tag{82}$$

unde

$$\tilde{\alpha}_i = n_i^{-1}\sum_{j=1}^{n_i} Y_{ij} - \bar{c}_{i,N}\left(\sum_{j=1}^{n_i} Y_{ij}(c_{ij} - \bar{c}_{i,N})/C_{i,N}^2\right),\ 1 \leqslant i \leqslant k,$$

$$\tilde{\tilde{\alpha}} = \left\{\sum_{i=1}^{k} n_i\tilde{\alpha}_i/(1 + q_{i,N}^2)\right\}\Big/\left\{\sum_{i=1}^{k} n_i/(1+q_{i,N}^2)\right\} \text{ și}$$

s_e^2 este dispersia datorită erorii cu $N - 2k$ grade de libertate.

Sub $H_0(64)$, Z_N are repartiția raportului dispersiilor cu $k-1$, respectiv $N - 2k$ grade de libertate și testul bazat pe Z_N este cel mai puternic și invariant.

Cînd F nu mai este normală, optimalitatea testului bazat pe Z_N se pierde. Totuși dacă $\sigma^2(F)$, dispersia variabilei avînd funcția de repartiție F, este finită,

i) $s_e^2 \xrightarrow[N\to\infty]{P} \sigma^2(F)$,

(ii) $n_i^{1/2}(\tilde{\alpha}_i - \alpha_i)(1 + q_{i,N}^2)^{-1/2}/\sigma(F)$ are asimptotic o repartiție $N(0, 1)$, $1 \leqslant i \leqslant k$.

Deci, asimptotic sub $H_0(64)$

$$(k-1)\,Z_N \sim \chi^2(k-1).$$

Astfel Z_N dă un test cu repartiția liberă pentru întreaga clasă a funcțiilor F cu $\sigma^2(F) < \infty$.

Dacă considerăm șirul alternativelor Pitman $\{H_N\}$ specificate de

$$H_N;\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k) = \boldsymbol{\alpha}_N = \alpha_0\mathbf{1} + N^{-1/2}\boldsymbol{\theta},\quad \mathbf{1} = (1, \ldots, 1) \tag{83}$$

unde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \neq 0$ și presupunem că:

$$\lim_{N\to\infty} q_{i,N} = q_i \text{ există}, \tag{84}$$

atunci sub $\{H_N\}$, Z_N are asimptotic o repartiție χ^2 cu $k-1$ grade de libertate de parametru de necentricitate

$$\Delta_Z = \left[\sum_{i=1}^{k} \lambda^{(i)}(\theta_i - \bar{\theta})^2(1 + q_i^2)^{-1}\right]/\sigma^2(F), \tag{85}$$

Scanned with OKEN Scanner

unde

$$\bar{\theta} = \left[\sum_{i=1}^{k} \lambda^{(i)} \theta_i (1+q_i^2)^{-1}\right]\left[\sum_{i=1}^{k} \lambda^{(i)} (1+q_i^2)^{-1}\right]^{-1}. \tag{86}$$

Din rezultatele lui Wald (1943) urmează că pentru clasa F cu $I(F) < \infty$ testul local optim asimptotic (în sensul că este asimptotic local cel mai stringent și are cea mai bună putere medie pe o suprafață elipsoidală din spațiul parametrilor) pentru $H_0(64)$ este bazat pe testul raportului de verosimilitați l_N, unde $-2 \log l_N$ are asimptotic

i) sub $H_0(64)$, o repartiție $\chi^2(k-1)$,

ii) sub $\{H_N\}$ (83), o repartiție $\chi^2(k-1)$ și parametrul de necentricitate

$$\Delta_l = \left[\sum_{i=1}^{k} \lambda^{(i)} (\theta_i - \bar{\theta})^2 (1+q_i^2)^{-1}\right] I(F) = \Delta_Z [I(F)\sigma^2(F)] \geqslant \Delta_Z \tag{87}$$

(din inegalitatea lui Rao-Cramer)

Observație. Astfel, în general, testul Z_N nu este asimptotic optim.

1.2.3. Teste de rang de aliniere

Pentru aliniere, considerăm următorul estimator pentru $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$

Fie

$$\hat{\boldsymbol{\beta}}_{i,N}^{(1)} = \sup\{b : T_{i,N}(b) > 0\}, \quad \hat{\boldsymbol{\beta}}_{i,N}^{(2)} = \inf\{b : T_{i,N}(b) < 0\}; \tag{88}$$

$$\hat{\beta}_{i,N} = \frac{1}{2}(\hat{\beta}_{i,N}^{(1)} + \hat{\beta}_{i,N}^{(2)}), \quad 1 \leqslant i \leqslant k, \quad \hat{\boldsymbol{\beta}}_N = (\hat{\beta}_{1,N} \ldots, \hat{\beta}_{k,N}). \tag{89}$$

Atunci $\hat{\boldsymbol{\beta}}_N$ este un estimator robust, invariant la translație și consistent pentru $\boldsymbol{\beta}$. Considerăm statistica de rang de aliniere

$$S_{i,N}(a, \hat{\beta}_{i,N}), \quad i = 1, \ldots, k. \tag{90}$$

unde pentru fiecare a real $(-\infty < a < \infty)$ fie

$$S_N^*(a) = \sum_{i=1}^{k} w_{i,N} S_{i,N}(a, \hat{\beta}_{i,N}) \tag{91}$$

$$w_{i,n} = [n_i/(1+q_{i,N}^2)]/\left[\sum_{i=1}^{k} n_i/(1+q_{i,N}^2)\right], \quad i = 1, \ldots, k. \tag{92}$$

Vom estima valoarea comună (ipotetică) a lui α_0 prin $\hat{\alpha}_N$, unde

$$\hat{\alpha}_N^{(1)} = \sup\{a : S_N^*(a) > 0\}, \quad \hat{\alpha}_N^{(2)} = \inf\{a : S_N^*(a) < 0\}; \tag{93}$$

$$\hat{\alpha}_N = \frac{1}{2}(\hat{\alpha}_N^{(1)} + \hat{\alpha}_N^{(2)}). \tag{94}$$

Scanned with OKEN Scanner

Testul va fi bazat pe statistica :

$$Q_N = \left[\sum_{i=1}^{k} n_i S_{i,N}^2(\hat{\alpha}_N, \hat{\beta}_{i,N})/(1 + q_{i,N}^2)\right] A^{-2}. \tag{95}$$

Au loc teoremele :

TEOREMA 8 (Sen (1972)). *Sub* (64) *şi presupunerile* § 1.2.1., Q_N *are asimptotic o repartiţei* $\chi^2(k-1)$.

TEOREMA 9 (Sen (1972)). *Sub presupunerile* § 1.2.1. *şi* (83)—(84), Q_N *are asimptotic o repartiţie* χ^2 *cu* $k-1$ *grade de libertate şi parametru de necentricitate*

$$\Delta_Q = [B^2(F, \varphi)/A^2]\left[\sum_{i=1}^{k} \lambda^{(i)}(\theta_i - \bar{\theta})^2/(1 + q_i^2)\right] \tag{96}$$

1.2.4. Eficienţa asimptotică relativă

Fie

$$\rho(\varphi, \varphi(u, f)) = [B(F, \varphi)]/\{AI(F)\}. \tag{97}$$

Din (81)

$$0 \leqslant \rho(\varphi, \varphi(u, f)) \leqslant 1 \text{ şi } \rho(\varphi, \varphi(u, f)) = 1 \text{ dacă } \varphi(u) = \varphi(u, f),\ 0 < u < 1.$$

Din (87) şi (96) urmează că eficienţa asimptotică a testului Q_N relativ la testul raportului de verosimilităţi este

$$e_{Q,l} = \Delta_Q/\Delta_l = \rho^2(\varphi, \varphi(u, f)) \tag{98}$$

şi deci, dacă $\varphi(u) \equiv \varphi(u, f)$, obţinem că asimptotic testul Q_N are aceleaşi proprietăţi de local optim ca testul raportului de verosimilităţi.

Din (85) şi (96) urmează că eficienţa asimptotică a testului Q_N relativ la testul raportului dispersiilor este egal cu

$$e_{Q,z} = \sigma^2(F)B^2(F, \varphi)/A^2. \tag{99}$$

Astfel pentru testul Q_N bazat pe scorurile normale, $e_{Q,z}$ este mărginită inferioar de 1, marginea inferioară fiind atinsă dacă F este normală.

De asemenea, pentru scorurile Wilcoxon, (99) este mărginită inferior de 0,864 pentru toţi F, în timp ce pentru mulţi F ne-normali, este mai mare decît unu.

1.3. Teste de rang bazate pe rangurile simultane ale tuturor observaţiilor pentru identitatea mai multor drepte de regresie

1.3.1. Notaţii şi preliminarii

Pentru fiecare $i = 1, \ldots, k$, fie Y_{ij}, $j = 1, \ldots, n_i$ variabile aleatoare independente cu funcţiile de repartiţie F_{ij} continui şi date de

$$P(Y_{ij} \leqslant y) = F_{ij}(y) = F(y - \alpha - \beta_i x_{ij}). \tag{100}$$

Scanned with OKEN Scanner

Nu se presupune cunoscută forma precisă a funcției $F(.)$. Aici x_{ij} sînt constante de regresie cunoscute, α este un parametru nedorit și β_i sînt cantitățile ce prezintă interes.

Ne propunem să verificăm ipoteza

$$H_0 : \beta_j = \beta \text{ (necunoscut)} \tag{101}$$

față de mulțimea alternativelor că $\beta_1, \ldots, \beta_k$ nu sînt toți egali.

Folosim notațiile

$$x_{i\cdot} = n_i^{-1} \sum_j x_{ij},\ C_{ni}^2 = \sum_j (x_{ij} - x_{i\cdot})^2,\ \gamma_{ni} = \{C_{ni}/C_n\} \tag{102}$$

$$i = 1, \ldots, k$$

$$C_n^2 = \sum_i C_{ni}^2,\ n = \sum_i n_i \tag{103}$$

unde în acest paragraf sumele se fac după j de la 1 pînă la n_i.

Vom scrie

$$c_{sj}^{(i)} = \begin{cases} \gamma_{ni}(x_{sj} - x_{s\cdot}), & s = 1, \ldots, i-1, i+1, \ldots, k \\ (\gamma_{ni} - 1)(x_{sj} - x_{s\cdot}), & s = i \end{cases} \tag{104}$$

astfel că

$$\bar{c}^{(i)} = n^{-1} \sum_s \sum_j c_{sj}^{(i)} = 0, \quad i = 1, \ldots, k \tag{105}$$

și

$$\sum_s \sum_j \{c_{sj}^{(i)} - \bar{c}^{(i)}\}^2 = (\gamma_{ni} - \gamma_{ni}^2) C_n^2 \leqslant C_n^2. \tag{106}$$

Presupunem că fiecare dintre C_{ni}^2 (și deci C_n^2) tinde la ∞ cu n astfel că:

$$\lim \gamma_{ni} = \gamma_i;\ 0 \leqslant \gamma_0 \leqslant \gamma_1, \ldots \gamma_k \leqslant (1 - \gamma_0) < 1, \tag{107}$$

unde $\gamma_0 < 1/k$.

Fie R_{ij} rangul lui Y_{ij} în rangurile combinate ale tuturor celor n observați.

Trebuie să estimăm β din (101). În acest scop fie

$$S_n(Y) = \sum_i \sum_j (x_{ij} - x_{i\cdot}) a_n(R_{ij}), \tag{108}$$

unde a_n este definit de (XXXVI.118).

Definim (Adichie (1967), (1974)) estimația lui β bazată pe (108) după cum urmează

$$\beta_n^* = \sup\{b : S_n(Y - bx) > 0\};\ \beta_n^{**} = \inf\{b : S_n(Y - bx) < 0\} \tag{109}$$

$$\hat{\beta}_n = \frac{1}{2}(\beta_n^* + \beta_n^{**}) \tag{110}$$

Scanned with OKEN Scanner

unde $S_n(Y - bx)$ este statistica (108) cînd observaţiile Y_{ij} sînt înlocuite prin $(Y_{ij} - bx_{ij})$. Estimaţia definită în (110) este consistentă în sensul dat în lema 1 ce urmează. Punem

$\hat{Y}_{ij} = (Y_{ij} - \hat{\beta}x_{ij})$ şi fie $\hat{R}_{ij}$ rangul lui $\hat{Y}_{ij}$. Definim

$$\hat{T}_{ni} = T_{ni}(\hat{Y}) = \sum_s \sum_j c_{sj}^{(i)} a_n(\hat{R}_{sj}) \tag{111}$$

şi statistica test pe care o vom folosi este

$$\hat{L}_n = \sum_i (\hat{T}_{ni}/AC_{ni})^2, \tag{112}$$

unde A^2 este definită în (XXXVI.121).

1.3.2. Repartiţia asimptotică a statisticii $\hat{L}_n$

Pentru a studia puterea asimptotică a testului bazat pe $\hat{L}_n$, trebuie să stabilim repartiţia limită nu numai sub ipoteza H_0, ci şi sub şirul de alternative

$$H_n: \beta_{ni} = \beta + (\theta_i/C_n), \quad |\theta_i| \leqslant M_2, \quad i = 1, \ldots, k. \tag{113}$$

Fie $Y_{ij}^0 = (Y_{ij} - \beta x_{ij})$ şi fie R_{ij}^0 rangul lui Y_{ij}^0. De asemenea fie T_{ni}^0 şi L_n^0 statisticile T_{ni} şi L_n asociate cu Y_{ij}^0. Găsirea repartiţiei limită a statisticii $\hat{L}_n$ depinde de următoarele leme:

LEMA 1. *Fie $\hat{\beta}_n$ definit ca în (110). Presupunem că funcţia generatoare de scor $\varphi(u)$ şi constantele de regresie x_{ij} satisfac*

i) *derivata de ordinul doi a funcţiei φ este mărginită*

ii) $\sup_y [(\mathrm{d}/\mathrm{d}y)\varphi\{F(y)\}]$ *este de asemenea mărginită*

iii) $\max_{i,j} |x_{ij}| \leqslant M \max |x_{ij} - x_{i.}|$, *$M$ nu depinde de n*

iv) $\{\max_{i,j} (x_{ij} - x_{i.})^2/C_n^2\} \to 0$.

Atunci (cînd $n \to \infty$)

(114) $|C_n(\hat{\beta}_n - \beta)|$ *este mărginit în fiecare din probabilităţile P_0 (sub (101)) şi P_n (sub (113)).*

LEMA 2. *Pentru fiecare i, în condiţiile lemei 1*

$$\{(\hat{T}_{ni} - T_{ni}^0)/c_n\} \to 0$$

în fiecare din probabilităţile P_0 şi P_n.

LEMA 3. *Dacă $\hat{U}_{ni} = (\hat{T}_{ni}/AC_{ni})$, atunci în condiţiile lemei 2, vectorul $\hat{\mathbf{U}}' = (\hat{U}_{n1}, \ldots, \hat{U}_{nk})$ este asimptotic $N(\mathbf{0}, \boldsymbol{\Sigma})$ sub P_0 şi $N(\boldsymbol{\nu}_n, \boldsymbol{\Sigma})$ sub P_n, unde $\boldsymbol{\nu}_n' = (\nu_{n1}, \ldots, \nu_{nk})$* cu

$$\nu_{ni} = \gamma_{ni}^{1/2}\{\theta_i - \sum_s (\gamma_{ns}\theta_s)\} \int \varphi_V' \{F(y)\}\mathrm{d}F(y)/A \tag{115}$$

$$\boldsymbol{\Sigma} = (\sigma_{is}); \quad \sigma_{is} = \{\delta_{is} - (\gamma_{ni}\gamma_{ns})^{1/2}\}, \tag{116}$$

unde δ este simbolul lui Kronecker.

Scanned with OKEN Scanner

TEOREMA 10. (Adichie (1974)). *Considerăm modelul (100) și presupunem că sînt satisfăcute condițiile lemei 2. Atunci*

$$\lim P_0(\hat{L}_n \leqslant y) = P(\chi^2(k-1) \leqslant y), \tag{117}$$

$$\lim P_n(\hat{L}_n \leqslant y) = P(\chi^2(k-1,\ \Delta_L) \leqslant y), \tag{118}$$

unde

$$\Delta_L = \sum_i \nu_{ni}^2 = \{\sum_i \gamma_{ni}\theta_i^2 - (\sum_i \gamma_{ni}\theta_i)^2\}\left[\int \varphi_v'\{F(y)\}\,dF(y)\right]^2/A^2. \tag{119}$$

În baza relației (117) un test asimptotic de prag ε respinge ipoteza (101) dacă $\hat{L}_n$ este mai mare decît ε punctul superior corespunzător unei repartiții $\chi^2(k-1)$.

1.3.3. Eficiența asimptotică

Metoda obișnuită de verificare a ipotezei (101) în modelul (100) este bazată pe estimațiile de C.M.M.P $\overline{\beta}_n$ și $\overline{\beta}_{ni}$ ale parametrilor β și β_i. În calcularea estimațiilor, este ușor de lucrat cu

$$x'_{ij} = (x_{ij} - x_{i\cdot})$$

în loc de datele originale x_{ij}. Cu această reparametrizare, statistica test devine

$$M_n = \sum_i C_{ni}^2(\overline{\beta}_{ni} - \overline{\beta}_n)^2/(k-1)s_e^2, \tag{120}$$

unde

$$\hat{\beta}_{ni} = \sum_j Y_{ij}(x_{ij} - x_{i\cdot})/C_{ni}^2;\ \overline{\beta}_n = \sum_i \gamma_{ni}\overline{\beta}_{ni}, \tag{121}$$

unde s_e^2 este dispersia datorită erorii.

Este cunoscut că pentru orice funcție de repartiție $F(y)$ pentru care

$$\sigma^2(F) = \left\{\int y^2 dF(y) - \left(\int y dF(y)\right)^2\right\} < \infty,$$

$$\lim P_0\{(k-1)M_n \leqslant y\} = P(\chi^2(k-1) \leqslant y). \tag{122}$$

$$\lim P_n\{(k-1)M_n \leqslant y\} = P(\chi^2(k-1,\ \Delta_M) \leqslant y), \tag{123}$$

$$\Delta_M = \left\{\sum_i \gamma_{ni}\theta_i^2 - \left(\sum_i \gamma_{ni}\theta_i\right)^2\right\}/\sigma^2(F). \tag{124}$$

Prin urmare eficiența testului $\hat{L}_n$ relativă la testul obișnuit bazat pe C.M.M.P. este

$$\Delta_L/\Delta_M = \sigma^2(F)\left[\int \varphi_v'\{F(y)\}dF(y)\right]^2/A^2. \tag{125}$$

Scanned with OKEN Scanner

Eficiența din (125) este aceeași cu cea din §1.

Observație. Pentru testele bazate pe scoruri de rang, dacă alegem funcția generatoare de scor $\varphi(u, f)$ dată de (XXXVI. 119), din presupunerile lemei 2, testul $\hat{L}_n$ bazat pe $\varphi(u, f)$ cu $\int \varphi^2(u, f)du < \infty$ este asimptotic cel mai puternic test de rang pentru ipoteza (101) în modelul (100).

§ 2. Teste de rang pentru subipoteze în regresia liniară generală

2.1. Statisticile test de rang și semn

Considerăm modelul liniar general

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{126}$$

unde $\mathbf{Y}$ este vectorul observațiilor independente $n \times 1$
$\mathbf{X}$ este o matrice $n \times p$ de constante cunoscute
$\boldsymbol{\beta}$ este vectorul parametrilor regresiei necunoscut $p \times 1$ astfel încît

$$M(\boldsymbol{\varepsilon}) = \mathbf{0}, \quad M(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \mathbf{I}_n, \quad \sigma > 0, \tag{127}$$

unde $\mathbf{I}_n$ este matricea identitate de ordin n.

Se convine să se scrie $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ astfel încît (126) poate fi pus sub forma:

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}, \tag{128}$$

unde $\mathbf{X}_1$ și $\mathbf{X}_2$ sînt de ordin $n \times k$, respectiv $n \times (p - k)$, în timp ce $\boldsymbol{\beta}_1$ și $\boldsymbol{\beta}_2$ sînt subvectori ai lui $\boldsymbol{\beta}$ de ordin $k \times 1$ și respectiv $(p - k) \times 1$. Dorim să verificăm ipoteza

$$H_0: \boldsymbol{\beta}_1 = \mathbf{0}, \quad \boldsymbol{\beta}_2 \text{ nespecificat}, \tag{129}$$

față de alternativa

$$\boldsymbol{\beta}_1 \neq \mathbf{0}.$$

Forma funcțională precisă a funcției de repartiție $F(y/\sigma)$ a componentelor lui $\boldsymbol{\varepsilon}$ nu trebuie să fie cunoscută, însă vom presupune (Adichie (1978)) că satisface următoarele presupuneri.

Presupunerea A. Funcția de repartiție F are densitatea f simetrică, absolut continuă, astfel încît informația lui Fisher este finită.

În cele ce urmează, ne vom ocupa cu șiruri de vectori de variabilele aleatoare $\{Y_n\}$ și de matricile $\{\mathbf{X}_n\}$ $n = 1, 2, \ldots$ nealeatoare. În timp ce toate limitele sînt luate după $n \to \infty$, numărul p al parametrilor rămîne fixat. Vom folosi notația $\mathbf{X} = ((x_{ij})) = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, unde $\mathbf{x}_j$ reprezintă coloana j a matricei $\mathbf{X}$. De asemenea vom presupune că $\mathbf{X}$ satisface condițiile lui Kraft, van Eeden (1972):

Scanned with OKEN Scanner

Presupunerea B

i) $\{\max_i x_{ij}^2 / \sum_i x_{ij}^2\} \to 0$ pentru fiecare $j = 1, \ldots, p$

ii) rang $\mathbf{X}$, $r(\mathbf{X}) = p$,

iii) $n^{-1}(\mathbf{X}'\mathbf{X})$ tinde către matrice pozitiv definită $\mathbf{\Sigma} = ((\sigma'_{jj}))$,

iv) pentru fiecare pereche j, $k (j \neq k,\ j,\ k = 1, \ldots, p)$ există un număr $\gamma_{jk} \neq 0$, astfel încît pentru $n > n_0$,

a) $x_{ij}(x_{ij} + \gamma_{jk} x_{ik}) \geq 0$ pentru toţi i

b) $|\mathbf{x}_j|$ şi $|\mathbf{x}_j + \gamma_{jk}\mathbf{x}_k|$ sînt similar ordonate.

Doi vectori $\mathbf{u}$ şi $\mathbf{v}$ sînt ordonaţi similar dacă

$$(u_i - u_s)(v_i - v_s) \geq 0 \text{ pentru toţi } i,\ s. \tag{130}$$

Observaţie. Din cauza lui B(iv), modelul de regresie prezent nu acoperă întreaga clasă a modelelor de regresie de rang maxim care sînt tratate obişnuit prin M.C.M.M.P.

Fie $a_n(i)$ definit ca în (XXXVI. 118), φ satisfăcînd condiţiile:

Presupunerea C. $\varphi(u)$ se exprimă ca diferenţa dintre două funcţii de patrat integrabile monotone nenegative, astfel încît

$$\int \varphi^2(u)du > 0.$$

Definim $\varphi^+(u,f)$ ca în (XXXVI. 138)

De asemenea este necesară estimaţia $\hat{\boldsymbol{\beta}}_2$ a parametrului $\boldsymbol{\beta}_2$ nespecificat. Introducem notaţia

$$\|\boldsymbol{\beta}\| = (\boldsymbol{\beta}'\boldsymbol{\beta})^{1/2}$$

şi presupunem că estimaţia satisface următoarele două condiţii:

Presupunerea D

i) Termenul $n^{1/2}\|\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\|$ este $\mathbf{0}_p(1)$ cînd $n \to \infty$, unde p se referă la probabilitatea sub (129).

ii) Pentru toţi $\boldsymbol{\beta}_2$, $\hat{\boldsymbol{\beta}}_2(\mathbf{Y} - \mathbf{X}_2\boldsymbol{\beta}_2) = \hat{\boldsymbol{\beta}}_2(\mathbf{Y}) - \boldsymbol{\beta}_2$, unde $\hat{\boldsymbol{\beta}}_2(\mathbf{Y})$ reprezintă estimaţia calculată cu ajutorul lui $\mathbf{Y}$.

Observăm că estimaţia de C.M.M.P. calculată sub H_0 satisface presupunerea *D*.

Pentru fiecare $i = 1, \ldots, n$, fie

$$\mathbf{Y}_i(\hat{\boldsymbol{\beta}}_2) = \hat{Y}_i = (\mathbf{Y} - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2)_i, \tag{131}$$

unde membrul drept al lui (131) reprezintă componenta i a vectorului rezidual $\mathbf{Y} - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2$.

Definim vectorul cu n componente prin

$$\Psi(\hat{\boldsymbol{\beta}}_2) = \{a_n(\hat{R}_i)\operatorname{sgn} \hat{Y}_i,\quad i = 1, \ldots, n\}' \tag{132}$$

unde

$$\operatorname{sgn} y = \begin{cases} 1, & y > 0 \\ -1, & y < 0 \end{cases}$$

Scanned with OKEN Scanner

și $\hat{R}_i$ este rangul valorii asolute $|\hat{Y}_i|$ printre

$$|\hat{Y}_1|, \ldots, |\hat{Y}_n|.$$

Pentru fiecare $j = 1, \ldots, p$ fie

(133) $$s_j(\hat{\boldsymbol{\beta}}_2) = \hat{s}_j = \mathbf{x}_j'\boldsymbol{\psi}(\hat{\boldsymbol{\beta}}_2)/A,$$

unde

(134) $$A^2 = \int \varphi^2(u)du.$$

Vectorul statisticilor (133) se scrie

(135) $$\mathbf{S}(\hat{\boldsymbol{\beta}}_2) = \hat{\mathbf{S}} = (\hat{s}_1, \ldots, \hat{s}_n)' = \mathbf{X}'\boldsymbol{\psi}(\hat{\boldsymbol{\beta}}_2)/A$$

și fie $\hat{\mathbf{S}}' = (\hat{\mathbf{S}}_1', \hat{\mathbf{S}}_2')$ partiția sa astfel încît

(136) $$\hat{\mathbf{S}}_2 = (\hat{s}_{k+1}, \ldots, \hat{s}_p)' = \mathbf{X}_2'\boldsymbol{\psi}(\hat{\boldsymbol{\beta}}_2)/A,$$

statistica rangului semnului ce va fi considerată este

$$E(\hat{\boldsymbol{\beta}}_2) = \hat{E} = \hat{\mathbf{S}}'(\mathbf{X}'\mathbf{X})^{-1}\hat{\mathbf{S}} - \hat{\mathbf{S}}_2'(\mathbf{X}_2'\mathbf{X}_2)^{-1}\hat{\mathbf{S}} =$$

$$= \boldsymbol{\psi}'(\hat{\boldsymbol{\beta}})\mathbf{W}\boldsymbol{\psi}(\hat{\boldsymbol{\beta}})/A^2,$$

unde $\mathbf{W}$ este o matrice simetrică idempotentă de ordin $n \times n$ definită de

(138) $$\mathbf{W} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'.$$

Observăm că $\mathbf{W}$ este ortogonală cu $\mathbf{X}_2$ în sensul că

(139) $$\mathbf{W}\mathbf{X}_2 = \mathbf{0}.$$

În continuare:

(140) $$\mathbf{W}\mathbf{X}_1 = \{\mathbf{I}_n - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\}\mathbf{X}_1.$$

Proprietatea de ortogonalitate a lui $\mathbf{W}$ și $\mathbf{X}_2$ care este crucială în teoria repartiției C.M.M.P. va fi de asemenea foarte folositoare în teoria repartiției statisticilor de rang. Este de interes primar această ortogonalitate și motivează folosirea lui $\mathbf{W}$ ca funcție de pondere. Vom arăta că $\hat{E}$ este asimptotic o statistică cu repartiție liberă pentru verificarea ipotezei (129). Testul va respinge ipoteza dacă $\hat{E}$ este mare. Pentru a vedea performanțele asimptotice ale puterii lui $\hat{E}$ trebuie să găsim nu numai repartiția sa limită sub (129), ci și sub șirul de alternative Pitman:

(141) $$H_n: \boldsymbol{\beta}_1 = n^{-1/2}\mathbf{b}_1, \quad \|\mathbf{b}_1\| < C.$$

TEOREMA 11 (Adichie (1978)). *Sub presupunerile* $A - D$

$$\lim P_0(\hat{E} \leqslant y) = P(\chi^2(k) \leqslant y), \tag{142}$$

$$\lim P_n(\hat{E} \leqslant y) = P(\chi^2(k, \Delta_{\hat{E}}) \leqslant y), \tag{143}$$

unde parametrul de necentricitate $\Delta_{\hat{E}}$ *este dat de*

$$\Delta_{\hat{E}} = \lim n^{-1}\{\mathbf{b}_1'\mathbf{X}_1'\mathbf{W}\mathbf{X}_1\mathbf{b}_1\}\mathbf{K}_F^2(\varphi) \tag{144}$$

şi

$$K_F(\varphi) = \int \varphi(u)\varphi^+(u, f)\mathrm{d}u/A \tag{145}$$

în timp ce P_0 *şi* P_n *reprezintă probabilitatea sub* (129) *şi respectiv* (141).

2.2. Statistici test de rang

Statisticile de rang, deosebite de statisticile de rang şi semn din § 2.1., pot fi folosite de asemenea la construcţia unui test cînd matricea $\mathbf{X}$ satisface o mulţime de presupuneri specificate în B_1.

Vom prezenta în acest paragraf astfel de teste de rang.

Considerăm modelul

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{146}$$

unde

$\mathbf{Y}$ este vectorul $n \times 1$ al observaţiilor independente

$\mathbf{X}$ este o matrice $n \times p_1$

$\boldsymbol{\theta}$ este vectorul parametrilor de regresie necunoscuţi $(p_1 \times 1)$ şi

$\boldsymbol{\varepsilon}$ satisface (127).

Rescriem (146) ca

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\theta}_1 + \mathbf{X}_2\boldsymbol{\theta}_2 + \boldsymbol{\varepsilon} \tag{147}$$

unde $\mathbf{X}_1$ şi $\mathbf{X}_2$ sînt de ordin $(n \times k)$ şi respectiv $n \times (p_1 - k)$, în timp ce $\boldsymbol{\theta}_1$ şi $\boldsymbol{\theta}_2$ sînt subvectori ai lui $\boldsymbol{\theta}$.

Se pune problema verificării ipotezei:

$$H_0 : \boldsymbol{\theta}_1 = \mathbf{0}, \quad \boldsymbol{\theta}_2 \text{ nespecificat} \tag{148}$$

faţă de alternativa $\boldsymbol{\theta}_1 \neq \mathbf{0}$.

Funcţia de repartiţie $F(y/\sigma)$ comună componentelor lui $\boldsymbol{\varepsilon}$ va satisface

Presupunerea A_1. Funcţia de repartiţie F are densitatea f care este absolut continuă astfel incît informaţia lui Fisher $I(f)$ este finită. Pentru matricea $\mathbf{X}$ fie

$$\mathbf{Z} = \mathbf{X} - \hat{\mathbf{X}} = ((x_{ij} - \bar{x}_j)) = ((z_{ij})),$$

Scanned with OKEN Scanner

unde $\bar{x}_j = n^{-1} \sum_i x_{ij}$ și fie $\mathbf{Z}=(\mathbf{Z}_1, \mathbf{Z}_2)$ corespunzînd la $\mathbf{X}=(\mathbf{X}_1, \mathbf{X}_2)$. Fie $\mathbf{Z} = (z_1, \ldots, z_{p_1})$ și presupunînd că $\mathbf{X}$ este astfel încît $\mathbf{Z}$ satisface condițiile lui Kraft și Eeden (1972) și anume:

Presupunerea B_1

i) $\max_i \{z_{ij}^2 / \sum_i z_{ij}^2\} \to 0$ pentru fiecare $j = 1, \ldots, p_1$;

ii) rang lui $\mathbf{Z}$, $r(\mathbf{Z}) = p_1$;

iii) $n^{-1}(\mathbf{Z}'\mathbf{Z})$ tinde către matricea pozitiv definită $\mathbf{\Sigma}^* = ((\sigma_{jj}^{*\prime}))$;

iv) pentru fiecare pereche j, $k (j \neq k,\ j, k = 1, \ldots, p_1)$ există un număr $\gamma_{jk} \neq 0$ astfel încît pentru $n > n_0$, z_j și $z_j + \gamma_{jk} z_k$ sînt similar ordonate.

Observații. 1) Deoarece $r(\mathbf{X}) \leqslant r(\mathbf{Z}) + r(\bar{\mathbf{X}})$ și $r(\mathbf{X}) \leqslant p_1$, B_1(ii) va fi satisfăcută numai pentru $\mathbf{X}$ din (146) pentru care $r(\mathbf{X}) < p_1 + 1$, adică pentru unii $\mathbf{X}$ cu rang maxim p_1 sau mai mic. Un caz particular pentru care B^2(ii) are loc este orice matrice ortogonală cu $\bar{x}_1 = \ldots = \bar{x}_{p_1}$. Pe de altă parte observăm că B(ii) din paragraful 1 are loc pentru toți $\mathbf{X}$ de rang maxim.

2. Din cauza lui B_1(ii), (iii) și (iv), metoda scorurilor de rang descrisă în acest paragraf nu poate fi folosită în toate modelele liniare de rang maxim cînd se folosește M.C.M.M.P.

3. Procedeele de testare din acest paragraf sînt de asemene adevărate în condițiile lui Jurečková (1971) asupra lui $\mathbf{X}$.

Reamintim că scorurile $a_n(i) = \varphi(i/(n+1))$, $\varphi(u)$ satisfăcînd

Presupunerea C_1. $\varphi(u)$ nu este constantă și este exprimabilă ca diferența dintre două funcții monotone de pătrat integrabile pe (0,1).

Cu $\varphi(u, f)$ definit ca în (XXXVI.119) și $A^2(\varphi)$ ca în (XXXVI.121) fie

$$K_F(\varphi) = \int \varphi(u)\varphi(u, f)\, du / A(\varphi). \tag{149}$$

Se cere ca estimația $\tilde{\boldsymbol{\theta}}_2$ a lui $\boldsymbol{\theta}_2$ să satisfacă următoarele două condiții:

Presupunerea D_1

i) Termenul $n^{1/2} \|\tilde{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2\|$ este $\mathbf{0}_p(1)$ cînd $n \to \infty$; unde p se referă la probabilitatea sub (148);

ii) Pentru toți $\boldsymbol{\theta}_2$, $\tilde{\boldsymbol{\theta}}_2(\mathbf{Y} - \mathbf{Z}_2\boldsymbol{\theta}_2) = \tilde{\boldsymbol{\theta}}_2(\mathbf{Y}) - \boldsymbol{\theta}_2$ unde $\tilde{\boldsymbol{\theta}}_2(\mathbf{Y})$ reprezintă estimația calculată din $\mathbf{Y}$.

Pentru fiecare i, luăm

$$Y_i(\tilde{\boldsymbol{\theta}}_2) = \tilde{Y}_i = (\mathbf{Y} - \mathbf{Z}_2\tilde{\boldsymbol{\theta}}_2)$$

și definim vectorul cu n componente

$$\Phi(\tilde{\boldsymbol{\theta}}_2) = \{a_n(\tilde{R}_i),\quad i = 1, \ldots, n\}'. \tag{150}$$

unde $\tilde{R}_i$ este rangul lui $\tilde{Y}_i$ în rangurile variabilelor $\tilde{Y}_1, \ldots, \tilde{Y}_n$. Observăm că vectorul

$$\{\tilde{R}_i,\ i = 1, \ldots, n\}$$

Scanned with OKEN Scanner

rămîne neschimbat dacă înlocuim $\tilde{\tilde{Y}}_i$ prin $(\tilde{Y} - \tilde{X}_2\tilde{\theta}_2)$, $i = 1, \ldots, n$. Scriem

$$S(\tilde{\theta}_2) = \tilde{S} = (\tilde{s}_1, \ldots, \tilde{s}_p)' = Z'\Phi(\tilde{\theta}_2)/A(\varphi) \tag{151}$$

unde

$$s_j(\tilde{\theta}_2) = \tilde{s_j} = z_j'\Phi(\tilde{\theta}_2)/A(\varphi), \quad j = 1, \ldots, p_1 \tag{152}$$

Din partiția $\tilde{S} = (\tilde{S}_1', \tilde{S}_2')$ și statisticile de rang propuse

$$E(\tilde{\theta}_2) = \tilde{E} = \tilde{S}'(Z'Z)^{-1}\tilde{S} - \tilde{S}_2'(Z_2'Z_2)^{-1}\tilde{S}_2 =$$

$$= \Phi'(\tilde{\theta}_2)\tilde{W}\Phi(\tilde{\theta}_2)/A^2(\varphi), \tag{153}$$

unde $\tilde{W}$ este o matrice de ordin $n \times n$ simetrică și idempotentă obținută din (138) înlocuind X prin Z. Folosind $E(\tilde{\theta})$ pentru verificarea problemei, ipoteza (148) este respinsă pentru valori mari ale lui $\tilde{E}$.

Dacă considerăm șirul de ipoteze alternative

$$H_n : \theta_1 = n^{-1/2}\theta_1, \quad \|\theta_1\| < c, \tag{154}$$

obținem

TEOREMA 12 (Adichie (1978)). *In presupunerile* $A_1 - D_1$,

$$\lim P_0(\tilde{E} \leqslant y) = P(\chi^2(k) \leqslant y) \tag{155}$$

$$\lim P_n(\tilde{E} \leqslant y) = P(\chi^2(k, \Delta_E^*) \leqslant y), \tag{156}$$

unde parametrul de necentricitate Δ_E^* *are expresia*

$$\Delta_E^* = \lim n^{-1}\{\theta_1'Z_1'\tilde{W}'Z_1\theta_1\}K_F^2(\varphi), \tag{157}$$

în timp ce P_0 *și* P_n *reprezintă probabilitatea sub* (148) *și respectiv* (154).

2.3. Eficiența asimptotică relativă

Dacă modelul dat în (126) și (127) este de rang maxim și dacă repartiția lui ε este normală, statistica folosită pentru verificarea ipotezei (129) este bazată pe raportul de verosimilitate maximă

$$Q = (n - p)D_1/k\,D_0, \tag{158}$$

unde

$$D_1 + D_0 = (Y - X_2\bar{\bar{\beta}}_2)'(Y - X_2\bar{\bar{\beta}}_2), \tag{159}$$

și

$$D_0 = (Y - X\tilde{\beta})'(Y - X\tilde{\beta}) = Y'\{I_n - X(X'X)^{-1}X'\}Y \tag{160}$$

Scanned with OKEN Scanner

cu $\bar{\bar{\boldsymbol{\beta}}}_2$ și $\tilde{\boldsymbol{\beta}}$ estimatorii de C.M.M.P. pentru $\boldsymbol{\beta}$ sub (129), respectiv (126). Cu aceasta, Q are repartiția raportului de dispersii cu $(k, n-p)$ grade de libertate și testul care respinge H_0 pentru valori mari ale lui Q este C.M.P. test invariant. Cînd ipoteza de normalitatea este înlăturată, Q pierde din optimalitate și repartiția sa exactă este necunoscută. Totuși pentru orice funcție de repartiție F marginală a componentelor lui $\boldsymbol{\varepsilon}$, pentru care dispersia $\sigma^2 = \sigma^2(F)$ este finită, se poate arăta că

$$|n^{-1}\mathbf{D}_0 - \sigma^2| = o_p(1) \quad \text{cînd} \quad n \to \infty.$$

Mai departe, luînd

$$\mathbf{Y}(\boldsymbol{\beta}_2) = \mathbf{Y} - \mathbf{X}_2\boldsymbol{\beta}_2,$$

$\mathbf{D}_1$ poate fi scris ca

$$\mathbf{D}_1 = \mathbf{Y}'\mathbf{W}\mathbf{Y} = \mathbf{Y}'(\boldsymbol{\beta}_2)\mathbf{W}\mathbf{Y}(\boldsymbol{\beta}_2) + \boldsymbol{\beta}_2'\mathbf{X}_2'\mathbf{W}\mathbf{X}_2\boldsymbol{\beta}_2 = \tag{139}$$

$$= \mathbf{Y}'(\boldsymbol{\beta}_2)\mathbf{W}\mathbf{Y}(\boldsymbol{\beta}_2) = \mathbf{Y}'(\boldsymbol{\beta})\mathbf{L}\mathbf{V}\mathbf{L}'\mathbf{Y}(\boldsymbol{\beta}_2),$$

unde $\mathbf{Y}'(\boldsymbol{\beta}_2)\mathbf{L} = (\sum_i l_{ij} Y_i(\boldsymbol{\beta}_2),\ j = 1, \ldots, p)$

$$\mathbf{V} = \mathbf{I}_p - \mathbf{D}\mathbf{D}', \quad \mathbf{D} = \mathbf{B}_2(\mathbf{B}_2'\mathbf{B}_2)^{-1/2}.$$

Ținînd seama că dacă $\mathbf{X}$ satisface B(i) și B(iii), atunci

$$\lim \{\max_i l_{ij}^2 / \sum_i l_{ij}^2\} = 0, \quad j = 1, \ldots, p \tag{161}$$

urmează că sub H_0 (asimptotic)

$$\mathbf{L}'\mathbf{Y}(\boldsymbol{\beta}_2)/\sigma \sim N(\mathbf{0}, \mathbf{I}_p)$$

și sub H_n(141), $\mathbf{L}'\mathbf{Y}(\boldsymbol{\beta}_2)/\sigma$ are media $n^{-1/2}\mathbf{L}'\mathbf{X}_1\mathbf{b}_1/\sigma$.

Prin urmare, are loc

TEOREMA 13 (Adichie (1978)). *Dacă în modelul* (126) *și* (127) *componentele lui* $\boldsymbol{\varepsilon}$ *au funcția de repartiție comună* $F(y/\sigma)$ *cu* $0 < \sigma < \infty$, *atunci*

$$\lim P_0(kQ \leqslant y) = P(\chi^2(k) \leqslant y)$$

$$\lim P_n(kQ \leqslant y) = P(\chi^2(k, \Delta_Q) \leqslant y),$$

unde P_0, P_n *și* $\chi^2(k, \Delta)$ *sînt definite în teorema* 12 *și*

$$\Delta_Q = \lim n^{-1}\{\mathbf{b}_1'\mathbf{X}_1'\mathbf{W}\mathbf{X}_1\mathbf{b}_1\}/\sigma^2. \tag{162}$$

Prin metoda convențională de măsurare a eficienței asimptotice relative a două statistici test cu repartiții χ^2 cu același număr de grade de libertate, din (144) și (162) urmează că eficiența asimptotică a lui $\hat{E}$ relativă la C.M.M.P. este

$$e_{\hat{E},Q} = \Delta_{\hat{E}}/\Delta_Q = \sigma^2 K_F^2(\varphi). \tag{163}$$

Scanned with OKEN Scanner

§ 3. Estimarea parametrului de locaţie după un test preliminar asupra regresiei

3.1. Cazul unidimensional

3.1.1. Introducere

Fie $Y_1, \ldots, Y_n$ variabile aleatoare independente cu funcţiile de repartiţie

$$F_i(x) = P\{Y_i \leqslant x\} = F(x - \theta - \beta t_i) \tag{164}$$

$$-\infty < x < \infty, \quad i = 1, \ldots, n$$

absolut continui, unde

$t_n = (t_1, \ldots, t_n)$ este un vector de constante cunoscute (nu toate egale),

θ şi β sînt parametrii necunoscuţi şi

forma funcţiei poate sau nu să fie cunoscută.

Mai întîi ne interesează estimarea lui θ. Dacă $\beta = 0$ atunci Y_i sînt identic repartizate cu locaţia θ; notăm estimaţia parametrului θ prin $\hat{\theta}_n$. Pe de altă parte, dacă β este necunoscut, estimaţia lui θ depinde de estimaţia lui β; o astfel de estimaţie este notată prin $\tilde{\theta}_n$. Cînd valoarea adevărată a lui β nu este specificată, însă este suspectată a fi apropiată de 0, se foloseşte un test de semnificaţie privind β:

dacă $H_0 : \beta = 0$ este acceptabilă estimatorul folosit este $\hat{\theta}_n$.

dacă $H_0 : \beta = 0$ nu este acceptabilă estimatorul folosit este θ_n^*

Un astfel de estimator după un test preliminar asupra regresiei este notat prin θ_n^* Obişnuit θ_n^* nu este nedeplasat.

3.1.2. Estimatorii locaţiei

Fie $\mathcal{F}$ clasa tuturor funcţiilor de repartiţie absolut continui simetrice faţă de 0 cu densităţi de repartiţie absolut continui (aproape peste tot) avînd informaţia lui Fisher finită.

Fie

$$\bar{t}_n = n^{-1}\sum_{i=1}^{n} t_i \text{ şi } Q_n = \sum_{i=1}^{n}(t_i - \bar{t}_n)^2. \tag{165}$$

Presupunînd că

(166) (164) are loc cu $F \in \mathcal{F}$

$$\lim_{n\to\infty} n^{-1} Q_n = Q^*(0 < Q^* < \infty) \text{ şi} \tag{167}$$

$$\lim_{n\to\infty} \bar{t}_n = \bar{t}(|\bar{t}| < \infty) \text{ există,}$$

(168) t_i sînt mărginiţi $(\Rightarrow \max_{1\leqslant i\leqslant n}(t_i - \bar{t})^2/Q_n \to 0$ cînd $n \to \infty)$

Scanned with OKEN Scanner

Fie $\varphi = \{\varphi(u),\ 0 < u < 1\}$ o funcţie strîmb-simetrică, adică

$$\varphi(u) + \varphi(1 - u) = 0 \quad (\forall)\ 0 < u < 1$$

şi de patrat integrabilă,

$$\varphi^* = \{\varphi^*(u) = \varphi((1 + u)/2),\ 0 < u < 1\}$$

şi pentru fiecare $n(\geqslant 1)$, fie

$$a_n(i) = M\varphi(U_{ni}) \text{ sau } \varphi\left(\frac{i}{n+1}\right),$$

(169)

$$a_n^*(i) = M\varphi^*(U_{ni}) \text{ sau } \varphi^*\left(\frac{i}{n+1}\right),\ i = 1, \ldots, n.$$

Fie $\mathbf{Y}_n = (Y_1, \ldots, Y_n)$ şi pentru fiecare (a, b) reali definim

$$\mathbf{Y}_n(a, b) = \mathbf{Y}_n - a\mathbf{1}_n - b\mathbf{t}_n,$$

unde $\mathbf{1}_n = (1, \ldots, 1)$.

Considerăm statisticile:

(170)
$$T_n(a, b) = T(\mathbf{Y}_n(a, b)) =$$

$$= n^{-1} \sum_{i=1}^{n} \operatorname{sgn}(Y_i - a - bt_i) a_n^*(R_{ni}^+(a, b)),$$

(171)
$$L_n(a, b) = L(\mathbf{Y}_n(a, b)) =$$

$$= n^{-1} \sum_{i=1}^{n} (t_i - \bar{t}_n) a_n(R_{ni}(a, b)),$$

unde $R_{ni}(a, b)(R_{ni}^+(a, b)))$ este rangul lui $Y_i - a - bt_i(|Y_i - a - bt_i|)$ printre $Y_1 - a - bt_1, \ldots, Y_n - a - bt_n(|Y_1 - a - bt_1|, \ldots, |Y_n - a - bt_n|)$, $i = 1, \ldots, n$.

Observăm că

$$R_{ni}(a, b) = R_{ni}(0, b)$$

pentru fiecare a real şi deci $L_n(a, b)$ nu depinde de a; scriem $L_n(b)$.

De asemenea, scriem

$$R_{ni}(0, 0) = R_{ni} \text{ pentru } i = 1, \ldots, n.$$

Pentru fiecare $\mathbf{Y}_n$ şi b daţi, $T_n(a, b)$ este $\downarrow$ în a: $-\infty < a < \infty$ şi pentru fiecare $\mathbf{Y}_n$, $L_n(b)$ este $\downarrow$ în b: $-\infty < b < \infty$ (teorema 7).

Scanned with OKEN Scanner

De asemenea dacă în modelul (164), luăm $\theta = \beta = 0$, atunci atît $T_n(0, 0)$ cît și $L_n(0)$ au repartiții simetrice față de zero. Considerăm următorii estimatori

(172) $$\hat{\theta}_n^{(1)} = \sup\{a : T_n(a, 0) > 0\}\ ;\ \hat{\theta}_n^{(2)} = \inf\{a : T_n(a, 0) < 0\}\ ;$$

(173) $$\hat{\theta}_n = \frac{1}{2}(\hat{\theta}_n^{(1)} + \hat{\theta}_n^{(2)})\ ;$$

(174) $$\tilde{\beta}_n^{(1)} = \sup\{b : L_n(b) > 0\}\ ;\ \tilde{\beta}_n^{(2)} = \inf\{b : L_n(b) < 0\}\ ;$$

(175) $$\tilde{\beta}_n = \frac{1}{2}(\tilde{\beta}_n^{(1)} + \tilde{\beta}_n^{(2)})\ ;$$

(176) $$\tilde{\theta}_n^{(1)} = \sup\{a : T_n(a, \tilde{\beta}_n) > 0\}, \tilde{\theta}_n^{(2)} = \inf\{a : T_n(a, \tilde{\beta}_n) < 0\}\ ;$$

(177) $$\tilde{\theta}_n = \frac{1}{2}(\tilde{\theta}_n^{(1)} + \tilde{\theta}_n^{(2)})$$

Atunci $\hat{\theta}_n$ este un estimator invariant la translație și consistent pentru θ (cînd $\beta = 0$) în timp ce $\tilde{\theta}_n$ are aceleași proprietăți cînd β nu este specificat.

Pentru un test preliminar asupra regresiei, folosim un test neparametric bazat pe $L_n = L_n(0)$. Astfel pentru verificarea ipotezei

$$H_0 : \beta = 0$$

față de

$$H_1 : \beta > 0$$

respingem H_0 dacă $L_n \geqslant L_{n,\alpha}$ unde

(178) $$P\{L_n \geqslant L_{n,\alpha} | H_0\} \leqslant \alpha, \quad 0 < \alpha < 1,$$

și α este pragul de semnificație al testului.

Dacă luăm

(179) $$A_n^2 = (n-1)^{-1} \sum_{i=1}^{n} \left\{a_n(i) - n^{-1} \sum_{j=1}^{n} a_n(j)\right\}^2 \quad , n \geqslant 2$$

și dacă τ_α este α-punctul superior al repartiției $N(0, 1)$, atunci

(180) $$n Q_n^{-1/2} \mathbf{A}_n^{-1} L_{n,\alpha} \to \tau_\alpha \text{ cînd } n \to \infty$$

Pentru valori mici ale lui n, $L_{n,\alpha}$ poate fi calculat prin enumerarea repartiției lui L_n sub H_0, generată de cele $n!$ permutări egal probabile ale rangurilor $R_{n1}, \ldots, R_{nn}$ (pe mulțimea $(1, \ldots, n)$ a numerelor întregi).

Atunci (Saleh, Sen (1978)) θ este estimat de

$$\theta_n^* = \begin{cases} \hat{\theta}_n, & \text{dacă } L_n < L_{n,\alpha} \\ \tilde{\theta}_n, & \text{dacă } L_n \geqslant L_{n,\alpha}. \end{cases}$$

Scanned with OKEN Scanner

Pentru un test bilateral, înlocuim în (178), (180) și (181), L_n prin $|L_n|$ și τ_α prin $\tau_{\alpha/2}$.
În cazul estimatorilor bazați pe teste preliminare, θ_n^* nu este un estimator nedeplasat. Se pune problema studierii naturii deplasării și a abaterii medii patratice pentru θ_n^*.

3.1.3. Repartiția asimptotică a estimatorului θ_n^*

Fie A_φ^2 (definit ca în (XXXVI.121), $\varphi^+(u, f)$ ca în (XXXVI.138)

$$A^2_{\varphi(u,f)} = I(F) = \int_0^1 \varphi^2(u, f)\, du$$

$$\gamma(\varphi, \varphi(u, f)) = \int_0^1 \varphi(u)\varphi(u, f) du. \tag{182}$$

Sub (167), (168) și (169) și $\theta = \beta = 0$ urmează că (Hájek, Šidák (1967)) pentru $n \to \infty$ vectorul $n^{1/2}(T_n, L_n)$, unde $T_n = T_n(0, 0)$ tinde în repartiție către un vector normal bidimensional

$$(\mathbf{0}, A_\varphi^2 \operatorname{diag}(1, Q^*) \tag{183}$$

astfel că dacă $\theta = \beta = 0$, cele două statistici sînt asimptotic independente).

Pentru $\beta = 0$, cînd $n \to \infty$

$$\sup\{n^{1/2}|L_n(n^{-1/2}b) - L_n(0) + \tag{184}$$

$$+ n^{-1/2} b\, Q^* \gamma(\varphi, \varphi(u, f))| \ |b| \leqslant K\} \xrightarrow{P} 0,$$

unde $K(0 < K < \infty)$ este o constantă pozitivă.

Dacă $\theta = \beta = 0$, Y_i și $-Y_i$ au aceeași densitate de repartiție.

De asemenea, după (168), există un $d(0 < d < \infty)$ astfel că $d|t_i| \leqslant 1$ pentru toți $i \geqslant 1$, așa că $(1 + dt_i) \geqslant 0$ pentru fiecare $i \geqslant 1$.
(167) și (168) asigură că pentru $\mathbf{x}_i = (1, t_i)$, $i \geqslant 1$,

$$n^{-1} \sum_{i=1}^n \mathbf{x}_i' \mathbf{x}_i \xrightarrow[n\to\infty]{} \begin{pmatrix} 1 & \bar{t} \\ \bar{t} & Q^* + \bar{t}^2 \end{pmatrix}$$

și de asemenea

$$\max_{1 \leqslant i \leqslant n} \left\{ t_i^2 \Big/ \sum_{i=1}^n t_i^2 \right\} \xrightarrow[n\to\infty]{} 0$$

Deoarece prima coordonată a lui $\mathbf{x}_i$ este egală cu 1 pentru toți i, urmează că

$$(|1| - |1|)(|1 + dt_i| - |1 + dt_{i'}|) = 0$$

pentru toți $i, i' = 1, \ldots, n$.

Scanned with OKEN Scanner

Urmează că pentru $\theta = \beta = 0$

$$\sup\{n^{1/2} | T_n(n^{-1/2}(a, b) -$$

$$(185) \qquad - T_n(0,0) + n^{-1/2}(a + b\tilde{t}_n)\gamma(\varphi, \varphi(u,f))| : |a| \leqslant$$

$$\leqslant K, \quad |b| \leqslant K\} \xrightarrow[n\to\infty]{P} 0,$$

unde $K(0 < K < \infty)$ este o constantă pozitivă.

Din (172), (177), (183), (184) și (185) urmează că repartiția vectorului $(n^{1/2}(\tilde{\theta}_n - \theta, \hat{\beta} - \beta))$ tinde către repartiția normală

$$(186) \qquad N_2\left(\mathbf{0}; \{A_\varphi^2/\gamma^2(\varphi, \varphi(u,f))\}\begin{pmatrix} 1 + \tilde{t}^2/Q^* & -\tilde{t}/Q^* \\ -\tilde{t}/Q^* & 1/Q^* \end{pmatrix}\right)$$

și cînd $\beta = 0$

$$(187) \qquad (n^{1/2}(\hat{\theta}_n - \theta)) \xrightarrow{R} N_1(0; A_\varphi^2/\gamma^2(\varphi, \varphi(u,f))).$$

Testul unilateral bazat pe L_n în (178) este consistent față de $\beta > 0$, astfel că asimptotic

$$P\{\theta_n^* = \tilde{\theta}_n | \beta > 0\} \to 1.$$

Deci, după (186), pentru fiecare $\beta > 0$, cînd $n \to \infty$

$$(188) \qquad (n^{1/2}(\theta_n^* - \theta)) \xrightarrow{R} N_1(0; A_\varphi^2(1 + \tilde{t}^2/Q^*)/\gamma^2(\varphi, \varphi(u,f))).$$

Similar,

$$P\{L_n < L_{n,\alpha} | \beta < 0\} \xrightarrow[n\to\infty]{} 1 \text{ astfel că } P\{\theta_n^* = \tilde{\theta}_n | \beta < 0\} \xrightarrow[n\to\infty]{} 1.$$

Pe de altă parte, pentru $\beta \neq 0$, $\tilde{t} \neq 0$ și orice x real $(-\infty < x < \infty)$

$$(189) \qquad P\{n^{1/2}(\theta_n^* - \theta) \leqslant x | \beta < 0\} \to 0 \text{ sau } 1$$

după cum $\tilde{t}$ este $<$ sau > 0.

Pentru un test bilateral preliminar, (188) are loc pentru orice $\beta \neq 0$. În fiecare caz, pentru $\beta = 0$ sau apropiat de 0, repartiția va fi diferită. În acest scop considerăm șirul de alternative $\{K_n\}$, unde

$$(190) \qquad K_n : \beta - \beta_{(n)} = n^{-1/2}\lambda, \quad \lambda \text{ real}$$

Atunci avem următoarea teoremă

TEOREMA (14 (Saleh, Sen (1978)). *Cu* $\{K_n\}$ *din* (190), *cînd* $n \to \infty$

$$(191) \qquad (n^{1/2}(\tilde{\theta}_n - \theta, L_n)) \xrightarrow{R} N_2(\mathbf{0}, \lambda Q^*\gamma(\varphi, \varphi(u,f)); \Sigma_1),$$

$$(192) \qquad (n^{1/2}(\hat{\theta}_n - \theta, L_n)) \xrightarrow{R} N_2(\lambda(\tilde{t}, Q^*\gamma(\varphi, \varphi(u,f)); \Sigma_2),$$

Scanned with OKEN Scanner

unde

$$\Sigma_1 = A_\varphi^2 \begin{pmatrix} (1 + \bar{t}^2/Q^*)\gamma^2(\varphi,\ \varphi(u,f)) & -\bar{t}/\gamma(\varphi,\ \varphi(u,f)) \\ -\bar{t}/\gamma(\varphi,\ \varphi(u,f)) & Q^* \end{pmatrix}$$

$$\Sigma_2 = A_\varphi^2 \begin{pmatrix} 1/\gamma^2(\varphi,\ \varphi(u,f)) & 0 \\ 0 & Q^* \end{pmatrix}. \tag{193}$$

Fie P_{K_n} probabilitatea sub K_n în (190). Atunci după (181), obţinem că pentru fiecare x real $(-\infty < x < \infty)$

$$P_{K_n}\{n^{1/2}(\theta_n^* - \theta) \leqslant x\} = P_{K_n}\{n^{1/2}(\theta_n^* - \theta) \leqslant x,\ L_n < L_{n,\alpha}\} +$$

$$+ P_{K_n}\{n^{1/2}(\theta_n^* - \theta) \leqslant x,\ L_n \geqslant L_{n,\alpha}\} = \tag{194}$$

$$= P_{K_n}\{n^{1/2}(\hat{\theta}_n - \theta) \leqslant x,\ L_n < L_{n,\alpha}\} +$$

$$+ P_{K_n}\{n^{1/2}(\tilde{\theta}_n - \theta) \leqslant x,\ L_n \geqslant L_{n,\alpha}\}.$$

A_n^2 definit de (179) converge către A_φ^2 cînd $n \to \infty$. Deci, dacă notăm prin $G(x)$ (şi $g(x)$) funcţia (densitatea) de repartiţie a variabilei $N(0, 1)$, din (168), (180), (194) şi teorema 14, obţinem că sub $\{K_n\}$ din (190), cînd $n \to \infty$, pentru orice x real

$$P_{K_n}\{n^{1/2}(\theta_n^* - \theta)\gamma(\varphi,\ \varphi(u,f))/A_\varphi \leqslant x\} =$$

$$P_{K_n}\{n^{1/2}(\theta_n^* - \theta) \leqslant xA_\varphi/\gamma(\varphi,\ \varphi(u,\ f))\} \to \tag{195}$$

$$\to G(x - \lambda\nu_1)G(\tau_\alpha - \lambda\nu_2) + \int_{\tau_\alpha - \lambda\nu_2}^{\infty} G(x + w\nu_1/\nu_2)\mathrm{d}G(w) =$$

$$= G_1^*(x)\ (\textit{notaţie}),$$

unde

$$\nu_1 = \bar{t}\gamma(\varphi,\ \varphi(u,\ f))/A_\varphi, \tag{196}$$

$$\nu_2 = (Q^*)^{1/2}\gamma(\varphi,\ \varphi(u,\ f))/A_\varphi \text{ şi } \nu_1/\nu_2 = \bar{t}/(Q^*)^{1/2}.$$

Similar, pentru un test bilateral preliminar, repartiţia asimptotică este dată de :

$$G_2^*(x) = G(x - \lambda\nu_1)\{G(\tau_{\alpha/2} - \lambda\nu_2) - G(-\tau_{\alpha/2} - \lambda\nu_2)\} +$$

$$+ \left\{\int_{-\infty}^{-\tau_{\alpha/2} - \lambda\nu_2} + \int_{\tau_{\alpha/2} - \lambda\nu_2}^{\infty} G(x + w\nu_1/\nu_2)\mathrm{d}G(w)\right\}. \tag{197}$$

Atît G_1^* cît şi G_2^* depinde de α, λ, $\bar{t}$, Q^*, φ şi $\varphi(u, f)$. Astfel avem

Scanned with OKEN Scanner

TEOREMA 15 (Saleh, Sen (1978)). *Pentru teste uni şi bilaterale preliminare pentru* β, *sub* (190), *repartiţiile asimptotice ale statisticii*

$$n^{1/2}(\theta_n^* - \theta)\gamma(\varphi, \varphi(u, f))/A_\varphi$$

sînt G_1^* *şi respectiv* G_2^*, *definite de* (195) *şi* (197).

Densităţile corespunzătoare funcţiilor de repartiţie G_1^* şi G_2^* sînt

$$g_1^*(x) = g(x - \lambda\nu_1)G(\tau_\alpha - \lambda\nu_2) + \int_{\tau_\alpha - \lambda\nu_2}^{\infty} g(x + w\nu_1/\nu_2)\,dG(w), \tag{198}$$

şi respectiv

$$g_2^*(x) = g(x - \lambda\nu_1)\{G(\tau_{\alpha/2} - \lambda\nu_2) - G(-\tau_{\alpha/2} - \lambda\nu_2)\} + \left\{\int_{-\infty}^{-\tau_{\alpha/2}-\lambda\nu_2} + \int_{\tau_{\alpha/2}-\lambda\nu_2}^{\infty} g(x + w\nu_1/\nu_2)\,dG(w)\right\}, \quad -\infty < x < \infty. \tag{199}$$

3.2. Cazul multidimensional

3.2.1. Introducere

Fie $\{\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})', i \geq 1\}$ un şir de vectori aleatori independenţi cu funcţiile de repartiţie $\{F_i, i \geq 1\}$, unde pentru fiecare $i \geq 1$,

$$F_i(\mathbf{x}) = F(\mathbf{x} - \boldsymbol{\theta} - \boldsymbol{\beta}t_i), \tag{200}$$

$$\mathbf{x} \in E^p, \ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)', \ \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$$

F, $\boldsymbol{\theta}$ şi $\boldsymbol{\beta}$ sînt necunoscuţi şi $\{t_i \geq 1\}$ este un şir de constante cunoscute. Sîntem interesaţi în primul rînd, în estimarea parametrului de locaţie θ.

Estimatorul parametrului θ este diferit după cum β este specificat sau nu. Fie $\hat{\boldsymbol{\theta}}_n$ şi respectiv $\tilde{\boldsymbol{\theta}}_n$ estimatorul parametrului $\boldsymbol{\theta}$ bazat pe $X_1, \ldots, X_n$, cînd presupunem că $\boldsymbol{\beta}$ este egal cu $\mathbf{0}$ şi cînd $\boldsymbol{\beta}$ nu este specificat. Obişnuit, $\tilde{\boldsymbol{\theta}}_n$ are o dispersie mai mare decît $\hat{\boldsymbol{\theta}}_n$ cînd $\boldsymbol{\beta} = \mathbf{0}$, în timp ce, pentru $\boldsymbol{\beta} \neq \mathbf{0}$, $\hat{\boldsymbol{\theta}}_n$ poate să nu fie un estimator convenabil pentru $\boldsymbol{\theta}$.
Cînd $\boldsymbol{\beta}$ este necunoscut însă este suspectat a fi apropiat de $\mathbf{0}$, adeseori, se foloseşte un test preliminat pentru $\boldsymbol{\beta}$:
dacă H_0 $\boldsymbol{\beta} = \mathbf{0}$ este acceptabilă, atunci este folosit $\hat{\boldsymbol{\theta}}_n$, în timp ce
dacă H_0 este respinsă, se foloseşte $\tilde{\boldsymbol{\theta}}_n$.

3.2.2. Notaţii preliminarii şi estimatori

Fie $R_{ni}^{(j)}(a, b)$ ($R_{ni}^{(j)*}(a, b)$) rangul lui $X_{ij} - a - bt_i$ ($|X_{ij} - a - bt_i|$) printre $X_{1j} - a - bt_1, \ldots, X_{nj} - a - bt_n$ ($|X_{1j} - a - bt_1|, \ldots, |X_{nj} - a - bt_n|$) pentru $i = 1, \ldots, n$, $j = 1, \ldots, p$, unde a şi b sînt numere reale.

Scanned with OKEN Scanner

Fie

$$\mathbf{T}_n(\mathbf{a}, \mathbf{b}) = (T_{n,1}(a_1, b_1), \ldots, T_{n,p}(a_p, b_p))' \quad \text{și}$$

$$\mathbf{S}_n(\mathbf{a}, \mathbf{b}) = (S_{n,1}(a_1, b_1), \ldots, S_{n,p}(a_p, b_p))',$$

unde, pentru fiecare $j = 1, \ldots, p$,

$$T_{n,j}(a, b) = n^{-1} \sum_{i=1}^{n} \operatorname{sgn}(X_{ij} - a - bt_i) a^*_{n,j}(R^{(j)*}_{ni}(a, b)) \tag{201}$$

$$S_{n,j}(a, b) = n^{-1} \sum_{i=1}^{n} (t_i - \bar{t}) a_{n,j}(R^{(j)}_{ni}(a, b)), \tag{202}$$

și scorurile $a_{n,j}(i)$ și $a^*_{n,j}(i)$ sînt definite ca în (169) cu ajutorul funcțiilor φ_j și respectiv φ_j^* și cu aceleași proprietăți, la care mai adăugăm ca φ_j este nedescrescătoare.

Presupunem că toți t_i sînt mărginiți și definim

$$Q_n^* = Q_n/n.$$

Presupunem că

$$\lim_{n\to\infty} \bar{t}_n = \bar{t}(|t| < \infty) \quad \text{și} \quad \lim_{n\to\infty} Q_n^* = Q^*(0 < Q^* < \infty) \quad \text{există.} \tag{203}$$

Fie $\mathcal{F}_p$ clasa densităților de repartiție p-dimensionale absolut continui care sînt simetric diagonale față de 0 și are matricea de informație a lui Fisher finită. Atunci presupunem că F din (200) aparține clasei $\mathcal{F}_p$.

Notăm că $T_{n,j}(a, b)$ este $\downarrow$ în a pentru (un b dat) și $S_{n,j}$ (a, b) este $\downarrow$ în b (independent de a) pentru fiecare $j = 1, \ldots, p$.

De asemenea sub

$$H_0^* : \boldsymbol{\theta} = \boldsymbol{\beta} = \mathbf{0},$$

$\mathbf{T}_n(\mathbf{0}, \mathbf{0})$ și $\mathbf{S}_n(\mathbf{0}, \mathbf{0})$ au media 0. Fie [Sen, Puri (1969), Sen, Saleh (1979)]

$$\hat{\theta}^{(1)}_{n,j} = \sup\{a : T_{n,j}(a, 0) > 0\}, \quad \hat{\theta}^{(2)}_{n,j} = \inf\{a : T_{n,j}(a, 0) < 0\} \tag{204}$$

$$\hat{\theta}_{n,j} = \frac{1}{2}(\hat{\theta}^{(1)}_{n,j} + \hat{\theta}^{(2)}_{n,j}); \; j = 1, \ldots, p; \; \hat{\boldsymbol{\theta}}_n = (\hat{\theta}_{n1}, \ldots, \hat{\theta}_{n,p})' \tag{205}$$

$$\hat{\beta}^{(1)}_{n,j} = \sup\{b : S_{n,j}(0, b) > 0, \quad \hat{\beta}^{(2)}_{n,j} = \inf\{b : S_{n,j}(0, b) < 0\} \tag{206}$$

$$\hat{\beta}_{n,j} = \frac{1}{2}(\hat{\beta}^{(1)}_{n,j} + \hat{\beta}^{(2)}_{n,j}); \; j = 1, \ldots, p; \; \hat{\boldsymbol{\beta}}_n = (\hat{\beta}_{n,1}, \ldots, \hat{\beta}_{n,p}) \tag{207}$$

$$\tilde{\theta}^{(1)}_{n,j} = \sup\{a : T_{n,j}(a, \hat{\beta}_{n,j}) > 0\}, \quad \tilde{\theta}^{(2)}_{n,j} = \inf\{a : T_{n,j}(a, \hat{\beta}_{n,j}) < 0\} \tag{208}$$

$$\tilde{\theta}_{n,j} = \frac{1}{2}(\tilde{\theta}^{(1)}_{n,j} + \tilde{\theta}^{(2)}_{n,j}), \quad j = 1, \ldots, p; \quad \tilde{\boldsymbol{\theta}}_n = (\tilde{\theta}_{n,1}, \ldots, \tilde{\theta}_{n,p})' \tag{209}$$

Scanned with OKEN Scanner

Atunci $\hat{\boldsymbol{\theta}}_n(\tilde{\boldsymbol{\theta}}_n)$ este un estimator invariant la translație, robust și consistent pentru $\boldsymbol{\theta}$ cînd $\boldsymbol{\beta} = \mathbf{0}$ (este nespecificat). Pentru un test preliminar pentru estimarea lui $\boldsymbol{\theta}_n^*$, fie

$$R_{ni}^{(j)} = R_{ni}^{(j)}(0, 0), \quad j = 1, \ldots, p; \; i = 1, \ldots, n$$

și definim $\mathbf{M} = ((m_{jl}))_{j,l=1,\ldots,p}$ prin

$$m_{jl} = n^{-1} \sum_{i=1}^{n} [a_{n,j}(R_{ni}^{(j)}) - \bar{a}_{n,j}][a_{n,l}(R_{ni}^{(l)}) - \bar{a}_{n,l}] \quad j, l = 1, \ldots, p \tag{210}$$

unde $\bar{a}_{n,j} = n^{-1} \sum_{i=1}^{n} a_{n,j}(i), \; j = 1, \ldots, p.$

Puri, Sen (1969), Sen, Saleh (1979) pentru verificarea ipotezei

$$H_0 : \boldsymbol{\beta} = \mathbf{0}$$

consideră statistica

$$\mathscr{L}_n = (n\mathbf{S}_n' \mathbf{M}_n^- \mathbf{S}_n)/Q_n^* \text{ unde } \mathbf{S}_n = \mathbf{S}_n(\mathbf{0}, \mathbf{0}) \tag{211}$$

și $\mathbf{M}_n^-$ este inversa generalizată a lui $\mathbf{M}$; $\mathscr{L}$ este o statistică avînd repartiția liberă și pentru valori mari ale lui n, sub H_0, ea are o repartiție $\chi^2(p)$ cînd $\mathbf{M}_n$ este de rang maxim. Notăm α punctul superior al repartiției statisticii $\mathscr{L}_n$ sub H_0 prin $\mathscr{L}_{n,\alpha}$ și notăm că

$$\mathscr{L}_{n,\alpha} \to \chi_\alpha^2(p)$$

α punctul superior al repartiției $\chi^2(p)$.

Atunci (Sen, Saleh (1979)) estimatorul pentru $\boldsymbol{\theta}$ după testul preliminar este

$$\boldsymbol{\theta}_n^* = \begin{cases} \hat{\boldsymbol{\theta}}_n, & \mathscr{L}_n \leqslant \mathscr{L}_{n,\alpha} \\ \tilde{\boldsymbol{\theta}}_n, & \mathscr{L}_n > \mathscr{L}_{n,\alpha} \end{cases} \tag{212}$$

Obișnuit, $\boldsymbol{\theta}_n^*$ nu este un estimator nedeplasat și nici asimptotic repartizat normal multidimensional. Din acest motiv, prezintă importanță studiul proprietăților asimptotice ale acestor estimatori.

3.2.3. Repartiția asimptotică a estimatorilor

Deoarece testul preliminar prezintă interes cînd $\boldsymbol{\beta}$ este suspect de a fi apropiat de $\mathbf{0}$, considerăm alternativele locale $\{K_n\}$ unde

$$K_n : \boldsymbol{\beta} = \boldsymbol{\beta}_{(n)} = n^{-1/2}\boldsymbol{\gamma}, \; \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)' \; (fixat) \in E^p \tag{213}$$

Scanned with OKEN Scanner

De asemenea, fie $F_{[j]}$ și respectiv $F_{[jl]}$ funcția de repartiție marginală a variabilei j și a repartiției comune a variabilelor (j, l), ce corespund la funcția de repartiție F (din (200) și fie

$$\lambda_{jl} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi_j(F_{[j]}(x))\,\varphi_l(F_{[l]}(y)\,dF_{[jl]}(x, y) \quad \text{pentru} \quad j, l = 1, \ldots, p \tag{214}$$

$$\mathbf{\Lambda} = ((\lambda_{jl})),\ \tau_{jl} = \lambda_{jl}/(\mu_j^*\,\mu_l^*) \quad \text{pentru} \quad j, l = 1, \ldots, p;\ \mathbf{T} = ((\tau_{jl}))$$

unde

$$\mu_j^* = \int_0^1 \varphi_j(u)\,\varphi_j(u, f_j)\,du$$

$$\varphi_j(u, f_j) = -f'_{[j]}(F^{-1}_{[j]}(u))/f_{[j]}(F^{-1}_{[j]}(u)),\ u \in (0, 1) \tag{215}$$

pentru $j = 1, \ldots, p$. Presupunem că atît $\mathbf{\Lambda}$ cît și $\mathbf{T}$ sînt de rang maxim. Atunci are loc

TEOREMA 16 (Sen, Saleh (1979)). *Sub $\{K_n\}$ din (213) și în condițiile de regularitate propuse*

$$n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \underset{R}{\rightarrow} \mathbf{N}_q(\bar{t}\boldsymbol{\gamma},\ \mathbf{T}) \tag{216}$$

și

$$n^{1/2}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \underset{R}{\rightarrow} N_p(\mathbf{0},\ (1 + \tilde{t}^2/Q^*)\mathbf{T}). \tag{217}$$

TEOREMA 17 (Sen, Saleh, (1979)). *Sub $\{K_N\}$ din (213) și în condițiile de regularitate ale teoremei 16, cînd $n \to \infty$*

$$\begin{aligned} G_p^*(\mathbf{x}, \boldsymbol{\gamma}) &= \lim_{n\to\infty} P_{K_n}\{n^{1/2}(\boldsymbol{\theta}_n^* - \boldsymbol{\theta}) \leqslant \mathbf{x}\} = \\ &= G_p(\mathbf{x} - \bar{t}\boldsymbol{\gamma};\ \mathbf{0},\ \mathbf{T})\,H_p(\chi_\alpha^2(p);\ \Delta^*) + \\ &+ \int_{E(\gamma)} G_p(\mathbf{x} + (\tilde{t}/Q^*)^{1/2}\,\mathbf{z};\ \mathbf{0}, \mathbf{T})\,dG_p(\mathbf{z}, \mathbf{0},\ T),\ (\forall)\mathbf{x} \in E^p \end{aligned} \tag{218}$$

unde $G_p(\mathbf{x};\ \boldsymbol{\mu}, \Sigma)$ este funcția de repartiție a unui vector p-dimensional cu vectorul medie $\boldsymbol{\mu}$ și matricea de covarianțe Σ, $H_p(x, \delta)$ este χ^2 necentrat cu p grade de libertate și parametru de necentricitate δ,

$$E(\gamma) = \{\mathbf{y} \in E^p : (\mathbf{y} + (Q^*)^{1/2}\boldsymbol{\gamma})'\,\mathbf{T}^{-1}(\mathbf{y} + (Q^*)^{1/2}\,\boldsymbol{\gamma}) > \chi_\alpha^2(p)\}$$

și

$$\Delta^* = Q^*(\boldsymbol{\gamma}'\mathbf{T}^{-1}\,\boldsymbol{\gamma}).$$

Dacă notăm prin g_p^* densitatea de repartiție corespunzătoare lui G_p^*, atunci pentru fiecare $\mathbf{x} \in E^p$,

$$\begin{aligned} g_p^*(\mathbf{x};\ \boldsymbol{\gamma}) &= g_p(\mathbf{x} - \bar{t}\boldsymbol{\gamma};\ \mathbf{0},\ \mathbf{T})\,H_p(\chi_\alpha^2(p);\ \Delta^*) + \\ &+ \int_{E(\gamma)} g_p(\mathbf{x} + (\bar{t}/Q^*)^{1/2}\,\mathbf{z};\ \ \mathbf{0},\ \mathbf{T})\,dG_p\,(\mathbf{z};\ \mathbf{0},\ \mathbf{T}) \end{aligned} \tag{219}$$

unde g_p este densitatea repartiției normale multidimensionale.

Scanned with OKEN Scanner

3.2.4. Deplasarea asimptotică şi matricea de covarianţe a estimatorilor

Vectorul medie şi matricea de covarianţe a funcţiei de repartiţie din (216), (217) şi (218) ale estimatorilor sînt definite ca *deplasarea asimptotică* şi *matricea de covarianţe asimptotică* a estimatorilor $\hat{\boldsymbol{\theta}}_n$, $\tilde{\boldsymbol{\theta}}_n$ şi respectiv $\boldsymbol{\theta}_n^*$. Astfel, avem

(220)
$$\xi_1(\boldsymbol{\gamma}) = \text{deplasarea asimptotică a lui } n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = \bar{t}\boldsymbol{\gamma}$$
$$\xi_2(\boldsymbol{\gamma}) = \text{deplasarea asimptotică a lui } n^{1/2}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = \mathbf{0}, (\forall)\boldsymbol{\gamma}$$
$$\boldsymbol{\Gamma}_1(\boldsymbol{\gamma}) = \text{matricea de covarianţe asimptotică } n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) =$$
$$= T + \bar{t}^2\boldsymbol{\gamma}\boldsymbol{\gamma}'$$

(221)
$$\boldsymbol{\Gamma}_2(\boldsymbol{\gamma}) = \text{matricea de covarianţe asimptotică } n^{1/2}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) =$$
$$= (1 + \bar{t}^2/Q^*)\,\mathbf{T}, (\forall)\boldsymbol{\gamma}.$$

Pentru a obţine expresii paralele pentru estimatorii după un test preliminar, notăm că prin definiţie pentru fiecare $c > 0$ şi $\delta = \mathbf{a}'\mathbf{B}^{-1}\mathbf{a}\,(\geqslant 0)$

(222)
$$\int_{\mathbf{x}'\mathbf{B}^{-1}\mathbf{x} > c} dG_p(\mathbf{x};\, \mathbf{a}, \mathbf{B}) =$$
$$= \left\{\exp\left(-\frac{1}{2}\delta\right)\right\}\left\{\sum_{r=0}^{\infty}\left(\frac{1}{2}\delta\right)^r [1 - H_{p+2r}(c\,;\,0)]/r\,!\right\}$$

şi deci prin derivare în raport cu $\mathbf{a}$, urmează că

(223)
$$\int_{(\mathbf{x}+\mathbf{a})'\mathbf{B}^{-1}(\mathbf{x}+\mathbf{a}) > c} \mathbf{x}\,dG_p(\mathbf{x}\,;\,\mathbf{0}, \mathbf{B}) = \mathbf{a}[H_p(c,\ \delta) - H_{p+2}(c,\ \delta)],$$

(224)
$$\int_{\mathbf{x}'\mathbf{B}^{-1}\mathbf{x} > c} (\mathbf{x} - \mathbf{a})(\mathbf{x} - \mathbf{a})'\, dG_p(\mathbf{x}\,;\ \mathbf{a},\ \mathbf{B}) = \int_{(\mathbf{x}+\mathbf{a})'\mathbf{B}^{-1}(\mathbf{x}+\mathbf{a}) > c} \mathbf{x}\mathbf{x}'dG_p\,(\mathbf{x}\,;\,\mathbf{0},\,\mathbf{B}) =$$
$$= [1 - H_{p+2}(c\,;\,\delta)\mathbf{B} - \mathbf{a}\mathbf{a}'\{H_p(c\,;\,\delta) - 2H_{p+2}(c\,;\,\delta) + H_{p+4}(c\,;\,\delta)\}].$$

Deci

$\boldsymbol{\xi}^*(\boldsymbol{\gamma})$ = deplasarea asimptotică a statisticii $n^{1/2}(\boldsymbol{\theta}_n^* - \boldsymbol{\theta}) =$

(225)
$$= \bar{t}\boldsymbol{\gamma}H_p(\chi_\alpha^2(p)\,;\,\Delta^*) + \int_{E(\gamma)} (-\bar{t}/(Q^*)^{1/2})\,\mathbf{z}\,dG_p(\mathbf{z}\,;\,\mathbf{0},\,\mathbf{T}) = \bar{t}\boldsymbol{\gamma}H_p(\chi_\alpha^2(p)\,;\,\Delta^*) +$$
$$+ (-\bar{t})(Q^*)^{1/2})(Q^*)^{1/2}\boldsymbol{\gamma}[H_p(\chi_\alpha^2(p)\,;\,\Delta^*) - H_{p+2}(\chi_\alpha^2(p)\,;\,\Delta^*)] =$$
$$= \{\bar{t}H_{p+2}(\chi_\alpha^2(p)\,;\,\Delta^*\}\,\boldsymbol{\gamma},$$

Scanned with OKEN Scanner

$$\Gamma^*(\gamma) = \text{matricea de covarianţe asimptotică a } n^{1/2}(\theta_n^* - \theta) =$$
$$= (1 + \bar{t}^2/Q^*)\mathbf{T} - (\bar{t}^2/Q^*)H_{p+2}(\chi_\alpha^2(p)\,;\,\Delta^*)\,\mathbf{T} -$$
(226)
$$- \bar{t}^2[H_p(\chi_\alpha^2(p)\,;\,\Delta^*) - 2H_{p+2}(\chi_\alpha^2(p)\,;\,\Delta^*) + H_{p+4}(\chi_\alpha^2(p)\,;\,\Delta^*)]\gamma\gamma'.$$

$\mathbf{T}$ care apare în (221) este pozitiv definită, în timp ce $\gamma\gamma'$ din (226) este semipozitiv definită (de rang 1), astfel că dacă $l_1 \geqslant \ldots \geqslant l_p$ sînt rădăcinile caracteristice ale lui $\mathbf{T}^{-1}\gamma\gamma'$, atunci $l_2 = \ldots = l_p = 0$ şi

(227) $$l_1 = \mathrm{Ur}[\mathbf{T}^{-1}\gamma\gamma'] = \gamma'\mathbf{T}^{-1}\gamma = \Delta^*/Q^*.$$

3.2.5. Compararea estimatorilor

Din (220), (221), (225) şi (226) urmează că pentru $\bar{t} = 0$,

(228) $$\xi_1(\gamma) = \xi_2(\gamma) = \xi^*(\gamma) = \mathbf{0} \text{ şi } \Gamma_1(\gamma) = \Gamma_2(\gamma) = \Gamma^*(\gamma) = \mathbf{T},\ (\forall)\gamma.$$

De asemenea, pentru $\bar{t} = 0$, după (218),

$$G_p^*(x\,;\,\gamma) = G_p(x\,;\,\mathbf{0},\,\mathbf{T}),\ (\forall)\gamma,$$

astfel că cei trei estimatori au asimptotic aceeaşi funcţie de repartiţie normală multidimensională; deplasarea lor asimptotică este egală cu $\mathbf{0}$ şi matricea de covarianţe asimptotică comună este $\mathbf{T}$. Pe de altă parte, pentru $\bar{t} \neq 0$, $\xi_2(\gamma) = \mathbf{0}$ şi $\Gamma_2(\gamma)$ nu depinde de γ în timp ce celelalte cantităţi depind.

Teorema 18 (Sen, Saleh (1979)). *Sub* $H_0 : \beta = \mathbf{0}$ (adică $\gamma = \mathbf{0}$), *cînd* $\bar{t} \neq \mathbf{0}$, $\Gamma_2(\mathbf{0}) - \Gamma_1(\mathbf{0})$, $\Gamma_2(\mathbf{0}) - \Gamma^*(\mathbf{0})$ *şi* $\Gamma^*(\mathbf{0}) - \Gamma_1(\mathbf{0})$ *sînt toate pozitiv definite, indiferent de* $\mathbf{T}$.

Demonstraţie. După (221), pentru $\gamma = \mathbf{0}$,

$$\Gamma_2(\mathbf{0}) - \Gamma_1(\mathbf{0}) = (\bar{t}^2/Q^*)\,\mathbf{T}$$

şi după (221) şi (226).

$$\Gamma_2(\mathbf{0}) - \Gamma^*(\mathbf{0}) = (\bar{t}^2/Q^*)H_{p+2}(\chi_\alpha^2(p)\,;\,0)\mathbf{T} \text{ şi}$$

$$\Gamma^*(\mathbf{0}) - \Gamma_1(\mathbf{0}) = (\bar{t}^2/Q^*)\{1 - H_{p+2}(\chi_\alpha^2(p)\,;\,0)\}\,T$$

unde

$$\bar{t}^2/Q^* > 0 \text{ şi } 0 < H_{p+2}(\chi_\alpha^2(p)\,;\,0) < H_p(\chi_\alpha^2(p)\,;\,0) = 1 - \alpha < 1.$$

Teorema este demonstrată.

Dacă definim eficienţa asimptotică relativă în termeni de dispersie generalizată asimptotică (adică rădăcina de ordinul p a determinantului matricei de covarianţe asimptotice), atunci pentru $\gamma = \mathbf{0}$

$$e(\theta^*, \hat{\theta}\,|\,\gamma = \mathbf{0}) = \{|\,\Gamma_1(\mathbf{0})\,|\,/\,|\Gamma^*(\mathbf{0})\,|\}^{1/p} =$$
(229)
$$= \{1 + (\bar{t}^2/Q^*)[1 - H_{p+2}(\chi_\alpha^2(p)\,;\,0)]\}^{-1}$$

Scanned with OKEN Scanner

unde $|\mathbf{B}|$ reprezintă determinantul lui $\mathbf{B}$ (= produsul rădăcinilor sale caracteristice).

Similar

$$e(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{\gamma} = \mathbf{0}) = \{|\boldsymbol{\Gamma}_2(\mathbf{0})| / |\boldsymbol{\Gamma}^*(\mathbf{0})|\}^{1/p} =$$

(230)

$$= (1 + \tilde{t}^2/Q^*)\{1 + (\tilde{t}^2/Q^*)[1 - H_{p+2}(\chi^2_\alpha(p)\,;\,0)]\}^{-1}.$$

Astfel

(231) $\quad e(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}} \,|\, \boldsymbol{\gamma} = \mathbf{0}) \leqslant 1 \leqslant e(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{\gamma} = \mathbf{0})$ pentru fiecare $\tilde{t}$,

unde, pentru $\tilde{t} = 0$, avem egalitate.

În cazul $\boldsymbol{\gamma} \neq \mathbf{0}$, din (221), (226) și (227) avem

$$e(\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{\gamma}) = \{|\boldsymbol{\Gamma}_2(\boldsymbol{\gamma})| / |\boldsymbol{\Gamma}_1(\boldsymbol{\gamma})|\}^{1/p} = (1 + \tilde{t}^2/Q^*)\{|\mathbf{I} + \tilde{t}^2\,\mathbf{T}^{-1}\,\boldsymbol{\gamma}\boldsymbol{\gamma}^{-1}|\}^{-1/p} =$$

(232)

$$= (1 + \tilde{t}^2/Q^*)\left\{\prod_{j=1}^{p}(1 + \tilde{t}^2 l_j)\right\}^{-1/p} = (1 + \tilde{t}^2/Q^*)(1 + \tilde{t}^2\, l_1)^{-1/p} =$$

$$= (1 + \tilde{t}^2/Q^*)(1 + \tilde{t}^2\Delta^*/Q^*)^{-1/p}.$$

Astfel (232) ia valori constante pe elipsoizii specificați de $\Delta^* = \boldsymbol{\gamma}'\mathbf{T}^{-1}\boldsymbol{\gamma}$ și cînd $\Delta^* \to 0$, tinde către $(1 + \tilde{t}^2/Q^*)$, în timp ce dacă $\Delta^* \to \infty$, tinde către 0, ceea ce indică că pentru valori mari ale lui Δ^*, $\hat{\boldsymbol{\theta}}_n$ devine relativ neeficient. În fapt

(233) $\quad e(\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{\gamma})$ este $\gtreqless 1$ după cum

$$\Delta^* \text{ este } \gtreqless [(1 + \tilde{t}^2/Q^*)^p - 1]/(\tilde{t}^2/Q^*).$$

§ 4. Convergența tare a statisticilor de rang de regresie

4.1. Introducere

Fie $\{X_i, i \geqslant 1\}$ un șir de variabile aleatoare independente definite pe spațiul măsurabil $(\Omega, \mathscr{A}, \mathscr{P})$ cu funcțiile de repartiție $\{F_i(x), i \geqslant 1\}$ continui.

Pentru o selecție $X_1, \ldots, X_n$ de volum n, fie

$$R_{ni} = \sum_{j=1}^{n} u(X_i - X_j)$$

rangul lui X_i printre $X_1, \ldots, X_n$ pentru $i = 1, \ldots, n$, unde $u(t)$ este definită ca în XXXVI § 4.1. Fie $\{c_i, i \geqslant 1\}$ un șir de constante de regresie

Scanned with OKEN Scanner

cunoscut. Definim (Hájek (1962, 1968) statistica de rang de regresie T_n prin

$$T_n = \sum_{i=1}^{c} (c_i - \bar{c}_n) J_n(R_{ni}/(n+1)) ; \ \bar{c}_n = n^{-1} \sum_{i=1}^{n} c_i, \ n \geqslant 1 \tag{234}$$

unde $c_1, \ldots, c_n$ nu sînt toţi egali

$$J_n(i/(n+1)) = M\varphi(U_{ni}), \ i = 1, \ldots, n$$

$\varphi(u)$, $0<u<1$ este funcţia scor ce satisface

$$\int_0^1 |\varphi(u)|^r \, du < \infty \text{ pentru } r > 2. \tag{235}$$

Fără a pierde din generalitate, luăm

$$\bar{\varphi} = \int_0^1 \varphi(u) \, du = 0$$

şi definim

$$A_n^2 = n^{-1} \sum_{i=1}^{n} J_n^2(i/(n+1)), \tag{236}$$

$$C_n^2 = \sum_{i=1}^{n} (c_i - \bar{c}_n)^2 \text{ şi } c_{ni}^* = (c_i - \bar{c}_n)/C_n, \ i = 1, \ldots, n \geqslant 1. \tag{237}$$

De asemenea presupunem (Hájek (1968)) că

$$\max_{1 \leqslant i \leqslant n} |c_{ni}^*| = \mathbf{0}(n^{-1/2}). \tag{238}$$

Fie

$$T_n^* = (n^{1/2} A_n C_n)^{-1} T_n = n^{-1/2} A_n^{-1} \sum_{i=1}^{n} c_{ni}^* J_n((n+1)^{-1} R_{ni}) ; \tag{239}$$

$$\bar{F}_n(x) = (n+1)^{-1} \sum_{i=1}^{n} F_i(x), \ F_n^*(x) = n^{-1/2} A_n^{-1} \sum_{i=1}^{n} c_{ni}^* F_i(x) ; \tag{240}$$

$$\tau_n^* = \int_{-\infty}^{\infty} \varphi(\bar{F}_n(x)) \, dF_n^*(x), \ n \geqslant 1. \tag{241}$$

Pentru fiecare întreg pozitiv, definim

$$Q(n) = C_n^2$$

Scanned with OKEN Scanner

și prin interpolare liniară, pentru $x \in (n, n+1)$ luăm

$$Q(x) = (x - n)\,Q(n+1) + (n + 1 - x)\,Q(n),\ n \geqslant 0,\ Q(0) = 0.$$

Atunci presupunem că constantele de regresie satisfac condițiile

(242) i) $Q(x)$ este $\uparrow$ în $x : 0 \leqslant x < \infty$, $\lim_{x\to\infty} Q(x) = \infty$;

(243) ii) pentru fiecare șir $\{a_n\}$ de numere pozitive pentru care

$$a_n \to 1 \text{ cînd } n \to \infty,\ Q(na_n)/Q(n) \to 1 \text{ cînd } n \to \infty$$

(244) iii) $\liminf_{n\to\infty} n^{-1} Q(n) \geqslant C_0^2 > 0$, $\limsup_{n\to\infty} n^{-h} Q(n) \leqslant C^* < \infty$,

unde $h(\geqslant 1)$ este un număr pozitiv. În final, fie $[s]$ cel mai mare întreg conținut în s.

(245) $$Q^{-1}(u) = \inf\{u : Q(x) \geqslant u\}$$

(246) $$V_n = \sum_{i=2}^{n} M\,[(T_i - T_{i-1})^2 \mid T_1, \ldots, T_{i-1}],\ n \geqslant 2$$

(247) $$\tilde{T}_{V_n} = T_n,\ n = 1, 2, \ldots,\ \tilde{T}_0 = T_0 = 0,$$

și prin interpolare liniară completăm definiția lui $\tilde{T}_t$ pentru fiecare $t \in (V_n, V_{n+1})$, $n \geqslant 1$. Atunci au loc următorele teoreme

TEOREMA 19 (Sen, Ghosh (1972)). *În presupunerile* (235), (238) *și*

(248) $$n^{-1} \sum_{i=1}^{n} |J_n(i/(n+1)) - \varphi(i/(n+1)| = \mathbf{0}(n^{-1/2}),$$

(249) $$\lim_{n\to\infty} (T_n^* - \tau_n^*) = 0 \text{ a.s.}$$

TEOREMA 20 (Sen, Glosh (1972)). *Dacă* $\varphi(u)$ *este nedescrescătoare în* u, (235), (238) *și* (244) *au loc,*

$F_t = F$ *pentru toți* $i \geqslant 1$ *și* $|\varphi'(u)| \leqslant K[u(1-u)]^{-3/2+\delta_0}$, $0 < u < 1$, $\delta_0 > 0$ *atunci*

a) *există o mișcare Browniană standard* $\xi(t)$ *pe* $[0, \infty)$ *astfel încît*

(250) $$\tilde{T}_t = \xi(t) + \mathbf{0}(t \log\log t)^{1/2} \text{ a.s. cînd } t \to \infty$$

Scanned with OKEN Scanner

și b)

$$\limsup_{n\to\infty} T_n/\{2A_n^2C_n^2 \log\log C_n^2\}^{1/2} = 1 \text{ a.s.}, \tag{251}$$

$$\liminf_{n\to\infty} T_n/\{2A_n^2C_n^2 \log\log C_n^2\}^{1/2} = -1 \text{ a.s.} \tag{252}$$

O aplicație a teoremelor 19 și 20 este următoarea. Luăm

$$F_i(x) = F(x - \beta_0 - \beta c_i),\ i \geqslant 1, \tag{253}$$

unde β este coeficientul de regresie, β este un parametru nedorit și c_i, $i \geqslant 1$ sînt constante de regresie cunoscute ce satisfac condițiile din acest paragraf. De asemenea presupunem că F este absolut continuă în raport cu măsura Lebesque, avînd o densitate $f(x)$ continuă, pozitivă și finită, astfel încît

$$\lim_{x\to\pm\infty} (\mathrm{d}/\mathrm{d}x)\varphi(F(x)) \text{ este finită.} \tag{254}$$

În final, presupunem că $\varphi(u)$ este strict crescătoare în $u : 0<u<1$, astfel încît

$$B(F) = \int_{-\infty}^{\infty} (\mathrm{d}/\mathrm{d}x)\varphi(F(x))\,\mathrm{d}F(x) > 0. \tag{255}$$

Pentru a verifica ipoteza

$$H_0 : \beta = 0 \text{ față de } H : \beta > 0 \tag{256}$$

poate fi formulat un procedeu asemănător cu cel din Darling, Robbins (1968). Fie

$$N = \begin{cases} \textit{primul întreg } n(\geqslant n_0) \textit{ pozitiv astfel încît } T_n > d_n \\ \infty, \textit{ dacă un astfel de } n \textit{ nu apare}, \end{cases} \tag{257}$$

unde $\{d_n\}$ este un șir de constante pozitive astfel încît $d_n \to \infty$ cînd $n \to \infty$. Testul constă în respingerea ipotezei H_0 dacă

$N<\infty$ *și acceptarea ipotezei* H_0 *în rest.*

Atunci teoremele 19 și 20 pot fi folosite pentru a arăta că acest test are puterea unu și volumul suficient de mic alegînd

i) n_0 suficient de mare și

ii) d_n crescînd cu o rată nu mai înceată decît a numărătorului din membrul drept al relației (251).

Fie $\mathscr{F}_n$ σ-algebra generată de $\mathbf{R}_n = (R_{n1}, \ldots, R_{nn})$: F_n este $\uparrow$ în $n(n \geqslant 1)$. În demonstrația teoremei 20 se folosește lema

LEMA 4. *Dacă* $F_1 = \ldots = F_n = F$ *pentru toți* $n \geqslant 1$, *atunci* $\{T_n, \mathcal{F}_n;$ $n \geqslant 1\}$ *este un șir martingal.*

Fie $C[0, 1]$ spațiul funcțiilor reale continue pe intervalul unitate $I = \{t : 0 \leqslant t \leqslant 1\}$. Pentru $n(\geqslant 2)$ definim procesul $Z_n = \{Z_n(t), \quad t \in I\}$ prin

$$Z_n(t_{k,n}) = A_n^{-1} C_n^{-1} T_k \tag{258}$$

unde

$$t_{k,n} = A_k^2 C_k^2 / A_n^2 C_n^2 \text{ pentru } k = 0, 1, \ldots, n;$$

și prin interpolare liniară, completăm definiția procesului $Z_n(t)$ pentru $t \in [t_{k,n}, t_{k+1,n}]$, $k = 0, 1, \ldots, n-1$. De asemenea introducem topologia uniformă

$$\rho(x, y) = \sup_{t \in I} |x(t) - y(t)|, \text{ unde } x, y \in C[0, 1]. \tag{259}$$

TEOREMA 21 (Sen, Ghosh (1972)). *În presupunerile teoremei* 20, $Z_n \to W$ *în topologia uniformă pe* $C[0, 1]$ *unde* $W = \{W(t), \quad t \in I\}$ *este o mișcare Browniană standard.*

4.2. Statisticele de regresie ale semnului rangului

Fie $\{X_i, i \geqslant 1\}$ un șir de variabile aleatoare independente cu funcțiile de repartiție $\{F_i(x), i \geqslant 1\}$, cu fiecare F_i absolut continuă în raport cu măsura Lebesque și simetrică față de zero, adică

$$F_i(x) + F_i(-x) = 1 \text{ pentru toți } x \text{ reali și } i \geqslant 1.$$

Fie

$$R_{ni}^+ = \sum_{j=0}^{n} u(|X_i| - |X_j|), \quad 1 \leqslant i \leqslant n \quad (\geqslant 1) \quad \text{și}$$

$$\tilde{T}_n = \sum_{i=1}^{n} c_i J_n^*(R_{ni}^+/(n+1)) \operatorname{sgn} X_i,$$

$$J_n^*(i/(n+1)) = M\varphi^*(U_{ni}) \text{ și } \varphi^*(u) = \varphi((1+u)/2), \quad 0 < u < 1$$

Presupunem că

$$\varphi(u) + \varphi(1-u) = 0 \text{ pentru toți } 0 < u < 1.$$

Scanned with OKEN Scanner

De asemenea fie

$$A_n^{*2} = n^{-1}\sum_{i=1}^{n}[J_n^*(i/(n+1))]^2,\quad \tilde{c}_{ni} = \left(\sum_{i=1}^{n} c_i^2\right)^{-1/2} c_i,\ i = 1, \ldots, n \text{ și}$$

$$\tilde{T}_n^* = n^{-1/2}A_n^{*-1}\sum_{i=1}^{n}\tilde{c}_{ni}\operatorname{sgn} X_i J_n^*(R_{ni}^+/(n+1)).$$

Cu acestea au loc teoremele

TEOREMA 22 (Sen, Ghosh (1972)). *Dacă* $\varphi \in L_r$ *pentru* $r > 2$ *și* $\max_{1\leq i\leq n}|\tilde{c}_{ni}| = \mathbf{0}(n^{-1/2})$, *atunci*

$$\lim_{n\to\infty}(\tilde{T}_n^* - \tau_n^*) = 0 \text{ a.s.}, \tag{260}$$

unde

$$\tau_n^* = n^{-1/2}A_n^{*-1}\sum_{i=1}^{n} c_{ni}\int_{-\infty}^{\infty}\operatorname{sgn} x\,\varphi^*(\bar{H}_n(|x|))\,\mathrm{d}F_i(x),$$

$$H_n(|x|) = n^{-1}\sum_{i=1}^{n} H_i(|x|) \text{ și}$$

$$\bar{H}_i(|x|) = F_i(|x|) - F_i(-|x|) = 2F_i(|x|) - 1,\ i \geq 1,\ x \text{ real}.$$

TEOREMA 23 (Sen, Ghosh (1972)). *Dacă*

i) $\varphi \in L_r$ *pentru* $r > 2$,

ii) *pentru fiecare* $0 < u < 1$,

$$|\varphi'(u)| < K[u(1-u)]^{-3/2+\delta_0} \text{ pentru } \delta_0 > 0,$$

iii) $$F_1 = \ldots = F_n = F\ (\forall) n > 1$$

iv) $$\max_{1\leq i\leq n}|\tilde{c}_{ni}| = \mathbf{0}(n^{-1/2}) \text{ și}$$

v) $$\liminf_{n\to\infty} n^{-1}\sum_{i=1}^{n} c_i^2 > C_0 > 0,\quad \limsup_{n\to\infty} n^{-h}\sum_{i=1}^{n} c_i^2 \leq C^* < \infty$$

atunci

a) *există o mișcare Browniană* $\xi(t)$ *pe* $[0, \infty)$ *astfel încît*

$$\tilde{T}_t^* = \xi(t) + o(t\log\log t)^{1/2}] \text{ a.s. } \textit{cînd } n \to \infty \tag{261}$$

Scanned with OKEN Scanner

unde

$$\tilde{T}^*_{V_n^*} = \tilde{T}_n, \quad \tilde{T}_0 = 0 \tag{262}$$

$$V_n^* = \sum_{t=1}^{n} M[(\tilde{T}_t - \tilde{T}_{t-1})^2 \mid \tilde{T}_1, \ldots, \tilde{T}_{t-1}] \tag{263}$$

și b)

$$\limsup_{n\to\infty} \tilde{T}_n[2(A_n^* C_n^*)^2 \log\log (A_n^* C_n^*)^2]^{-1/2} = 1 \text{ a.s}, \tag{264}$$

$$\liminf_{n\to\infty} \tilde{T}_n[2(A_n^* C_n^*)^2 \log\log (A_n^* C_n^*)^2]^{-1/2} = -1 \text{ a.s.} \tag{265}$$

unde

$$(C_n^*)^2 = \sum_{i=1}^{n} c_i^2. \tag{266}$$

Este posibil să înlocuim $(A_n^* C_n^*)^2$ prin V_n^*.

Acum considerăm un alt proces $Z_n^* = \{Z_n^*(t) : t \in I\}$, unde

$$Z_n^*(t_{k,n}) = (A_n^* C_n^*)^{-1} T_k^*,$$

$$t_{k,n} = (C_k^* A_k^*)^2/(C_n^* A_n^*)^2 : k = 0, 1, \ldots, n; \tag{267}$$

și prin interpolare liniară, completăm definiția lui $Z_n^*(t)$ pentru $t \in [t_{k,n}, t_{k+1,n}]$, $k = 0, 1, \ldots, n-1$. Are loc un analog al teoremei 21.

Teorema 24 (Sen, Ghosh (1972)). *În presupunerile teoremei* 23, $Z_n^* \to W$ *în topologia uniformă pe* $C[0, 1]$, *unde* $W = \{W(t) : t \in I\}$ *este mișcarea Browniană standard.*

Scanned with OKEN Scanner

Capitolul XXXVIII

REGRESIE ROBUSTĂ. PROCEDEE SECVENȚIALE

§ 1. Interval de încredere mărginit pentru coeficientul de regresie bazat pe o clasă de statistici de rang

1.1. Introducere

Considerăm modelul regresiei liniare

$$X_i = \beta_0 + \beta c_i + e_i, \quad i = 1, 2, \ldots \tag{1}$$

unde — c_i sînt constantele de regresie cunoscute și
e_i sînt variabile aleatoare independente și identic repartizate cu funcția de repartiție $F(x)$ (necunoscută) absolut continuă definită pe $(-\infty, \infty)$.

Dorim să obținem un interval de încredre robust (cu coeficientul de încredere $1 - \alpha$, $0<\alpha<1$) pentru coeficientul de regresie β, astfel încît lungimea acestui interval să fie mărginită de $2d$, pentru un $d>0$ specificat. Ghosh, Sen (1972) folosesc în acest scop clasa statisticilor de rang pentru regresie dată de Hájek (1962, 1968).

1.2. Notații și presupuneri

În concordanță cu modelul (1) considerăm șirul $\{X_1, X_2, \ldots\}$ de variabile aleatoare independente pentru care

$$F_i(x) = P\{X_i \leqslant x\} = F(x - \beta_0 - \beta c_i), \quad i = 1, 2, \ldots \tag{2}$$

unde β_0 este parametrul nedorit. Determinăm un interval de încredere

$$I_n = \{\beta : \hat{\beta}_{L,n} < \beta < \hat{\beta}_{U,n}\},$$

astfel încît

$$P\{\beta \in I_n\} = 1 - \alpha, \text{ coeficientul de încredere prestabilit} \tag{3}$$

$$0 < \hat{\beta}_{U,n} - \hat{\beta}_{L,n} < 2d, \text{ pentru un } d(>0) \text{ predeterminat.} \tag{4}$$

Scanned with OKEN Scanner

Folosim clasa de statistici de rang

$$T_n = T(X_n) = \sum_{i=1}^{n} c_{ni}^* J_n(R_{ni}/(n+1)), \tag{5}$$

unde c_{ni}^* şi R_{ni} au semnificaţiile din XXXVII. 4.1. şi scorurile $J_n(i/n+1))$, $1 \leqslant i \leqslant n$ sînt generate de „funcţia scor" $\{\varphi(u) : 0<u<1\}$ în următorul mod. Fie

$$J_n(u) = J_n\left(\frac{i}{n+1}\right) \text{pentru } (i-n)/n < u \leqslant i/n, \quad 1 \leqslant i \leqslant n \quad \text{şi}$$

$$J_n\left(\frac{i}{n+1}\right) \text{egal cu } \varphi\left(\frac{i}{n+1}\right) \text{sau } M\varphi(U_{ni}),\ 1 \leqslant i \leqslant n. \tag{6}$$

Funcţia scor $\varphi(u)$ este definită ca $\Psi^{-1}(u) : 0<u<1$, unde $\Psi(u)$ este o funcţie de repartiţie absolut continuă ce satisface condiţia

$$\lim_{u \downarrow 0} \psi(\Psi^{-1}(u))/u \geqslant K, \quad \lim_{u \uparrow 1} \psi(\Psi^{-1}(u))/(1-u) \geqslant K \tag{7}$$

unde $\psi(x) = (\mathrm{d}/\mathrm{d}x)\,\Psi(x)$ şi $0<K<\infty$.
(7) implică că

$$|\varphi(u)| \leqslant K_0[-\log u(1-u)],\ |\varphi'(u)| \leqslant K_0[u(1-u)]^{-1},\ 0<u<1,\ 0<K_0<\infty,$$

şi există un t_0 pozitiv astfel încît

$$G(t) = \int_{-\infty}^{\infty} \exp(tx)\,\mathrm{d}\Psi(x) < \infty \quad \text{pentru toţi} \quad 0<t \leqslant t_0 \tag{8}$$

Observaţii. i) Prin definiţie $\varphi(u)$ este $\nearrow$ în $u : 0<u<1$. Aceste presupuneri sînt mai restrictive decît acelea din Hájek (1962, 1968), cînd sînt studiate numai rezultate privind convergenţa slabă. Deoarece nu sîntem interesaţi în convergenţa tare pentru variabila de stopare ci şi în convergenţa în medie patratică, trebuie păstrate condiţiile anterioare.

ii) (7) are loc pentru repartiţiile : normală, dublu exponenţială, exponenţială, rectangulară şi pentru multe alte funcţii de repartiţie.

Asupra lui $\mathbf{c}_n = (c_1, \ldots, c_n)$ se fac următoarele presupuneri (Ghosh, Sen (1972))

$$\text{i) } \max_{1 \leqslant i \leqslant n} |c_{ni}^*| = \mathbf{0}(n^{-1/2}) \text{ (Hájek (1968))} \tag{9}$$

$$\text{ii) } \liminf_n n^{-1} C_n^2 \geqslant K_0 > 0, \tag{10}$$

iii) definim $Q(x) = (n+1-x)\,C_n^2 + (x-n)C_{n+1}^2$ pentru $n \leqslant x \leqslant n+1$, $n = 0, 1, \ldots,$ şi luăm $C_0^2 = 0$. Presupunem că $Q(x)$ este $\nearrow$ în x şi

$$\lim_{n \to \infty} Q(na_n)/Q(n) = s(a), \quad \lim_{n \to \infty} a_n = a \tag{11}$$

$s(a)$ fiind strict monotonă (crescătoare), cu $s(1) = 1$.

Scanned with OKEN Scanner

Observații. Condiția (9) este mai stringentă decît condiția clasică Noether (Hájek (1962), (1968)), însă este satisfăcută în majoritatea situațiilor practice.

(10) este mai puțin restrictivă decît condiția paralelă

$$\lim_{n\to\infty} n^{-1}C_n^2 = K_0 > 0$$

presupusă de Gleser (1965) și Albert (1966) în legătură cu teoria C.M.M.P. De exemplu, dacă $c_i = a + ih$, $h > 0$, $i = 1, 2, \ldots$, au loc (9) și (10) însă nu și condiția lui Gleser. De asemenea în acest caz $Q(n) = nh^2(n^2 - 1)/12$ astfel încît $s(a) = a^3$.

În final, presupunem că $F \in \mathcal{F}(\Psi)$, unde $\mathcal{F}(\Psi)$ este clasa tuturor funcțiilor de repartiție F pentru care densitatea $f(x)$ și prima sa derivată $f'(x)$ sînt mărginite pentru aproape toți x; de asemenea, presupunem că

(12) $$\lim_{x\to\pm\infty} f(x)\varphi'(F(x)) \text{ este finită.}$$

Din (2) urmează că sub H_0 : $\beta = 0$, $X_1, \ldots, X_n$ sînt variabile aleatoare independente, identic repartizate, $T(\mathbf{X}_n)$ are o repartiție complet specificată generată de cele $n!$ realizări ale lui $(R_{n1}, \ldots, R_{nn})$ egal posibile. Deci, există o constantă $T_n^{(1)}$ (ce depinde de C_n) și un α_n (cunoscut) astfel încît

(13) $$P\{-T_n^{(1)} \leqslant T(\mathbf{X}_n) \leqslant T_n^{(1)} \mid H_0\} = 1 - \alpha_n (\to 1 - \alpha \text{ cînd } n \to \infty)$$

(pentru valori mici ale lui n, α_n poate fi egal cu α). Fie

(14) $$A_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[J_n\left(\frac{i}{n+1}\right) - \bar{J}_n\right]^2; \quad \bar{J}_n = \frac{1}{n}\sum_{i=1}^{n} J_n\left(\frac{i}{n+1}\right);$$

atunci din teorema limită centrală de permutare (Wald-Wolfowitz-Noether-Hájek (Hájek, Sidák p. 160) avem

(15) $$\lim_{n\to\infty}\{T_n^{(1)}/A_n\} = \tau_{\alpha/2}$$

unde $\Phi(\tau_\varepsilon) = 1 - \varepsilon$.

Din XXXVII. 1.1.5 urmează că

$$T(\mathbf{X}_n - a\mathbf{c}_n) \text{ este } \nearrow \text{ în } a : -\infty < a < \infty.$$

Deci dacă luăm

(16) $$\hat{\beta}_{L,n} = \sup\{a : T(\mathbf{X}_n - a\mathbf{c}_n) > T_n^{(1)}\}, \quad \hat{\beta}_{U,n} = \inf\{a : T(\mathbf{X}_n - a\mathbf{c}_n) < -T_n^{(1)}\}$$

din XXXVII. 1.1.4 urmează că :

(17) $$P\{\hat{\beta}_{L,n} \leqslant \beta \leqslant \hat{\beta}_{U,n} \mid \beta\} = 1 - \alpha_n (\to 1 - \alpha \text{ cînd } n \to \infty).$$

Scanned with OKEN Scanner

Cu acestea Ghosh, Sen (1972) definesc următorul procedeu secvențial

Pentru fiecare $d>0$, fie $N(d)$ variabila de stopare cel mai mic întreg pozitiv $\geqslant n_0$ [volumul inițial al selecției ($\geqslant 3$)] pentru care

$$\hat{\beta}_{U,N(d)} - \hat{\beta}_{L,N(d)} \leqslant 2d.$$

Atunci intervalul de încredere pentru β este

$$I_{N(d)} = \{\beta : \hat{\beta}_{L,N(d)} \leqslant \beta < \hat{\beta}_{U,N(d)}\}$$

și este bazat pe regula de stopare $N(d)$. Acest interval este robust pentru date nesemnificative sau erori grosolane.

1.3. Comportarea asimptotică a unor procese empirice

Sub H_0: $\beta = 0$ pentru toți β reali și finiți, notînd

$W(\mathbf{X}_n, b) = T(\mathbf{X}_n) - T(\mathbf{X}_n - bc_n^*) - bB(F)$, pentru fiecare $c>0$ cu

$$B(F) = \int_{-\infty}^{\infty} (\mathrm{d}/\mathrm{d}x)\, \varphi[F(x)]\, \mathrm{d}F(x) \tag{18}$$

avem

$$\sup_{b:|b|<c} |\mathbf{W}(\mathbf{X}_n, b)| \xrightarrow{\mathbf{P}} 0 \quad \text{cînd} \quad n \to \infty. \tag{19}$$

Scopul nostru nu este numai de a întări această presupunere la convergența sigură, ci și a specifica ordinul convergenței, extinzînd domeniul lui b la

$$|b| \leqslant C(\log n)^k, \quad k \geqslant 1, \ 0 < C < \infty.$$

Definim

$$H_n(x, b) = \frac{1}{n+1} \sum_1^n u(x + bc_{ni}^* - X_i), \quad S_n^*(x, b) = \sum_1^n c_{ni}^* u(x + bc_{ni}^* - X_i), \tag{20}$$

pentru $-\infty < x < \infty$ și $-\infty < b < \infty$. Atunci după (5) și (20) avem :

$$T(\mathbf{X}_n - bc_n^*) = \int_{-\infty}^{\infty} J_n[H_n(x, b)]\, \mathrm{d}S_n^*(x, b). \tag{21}$$

Mai întîi vom studia comportarea asimptotică a celor două procese din (20), fapt ce ne va permite să dovedim (19) cu un ordin de convergență specificat.

Scanned with OKEN Scanner

Fie $\{Y_1, Y_2, \ldots\}$ un șir de variabile aleatoare independente, identic repartizate, avînd o repartiție rectangulară pe $(0, 1)$. Pentru fiecare $t : 0 < t < 1$ definim

$$d_{ni}(t\,;\,b) = F(F^{-1}(t) + bc_{ni}^*) - t, \quad 1 \leqslant i \leqslant n, \quad -\infty < b < \infty, \tag{22}$$

(23) i) $0 < t + d_{ni}(t\,;\,b) \leqslant 1$, $t + d_{ni}(t, b)$ este $\nearrow$ în $t : 0 < t < 1$,

(24) ii) $d_{ni}(t\,;\,b)$ și bc_{ni}^* au același semn,

$$\text{(25) iii)} \quad \bar{d}_n(t, b) = n^{-1} \sum_1^n d_{ni}(t, b) = b^2[\mathbf{0}(n^{-1})] \quad (\forall)\, F \in \mathscr{F}(\Psi).$$

Considerăm următoarele două procese stochastice :

$$G_n^*(t\,;\,b) = \sum_1^n c_{ni}^* u(t + d_{ni}(t\,;\,b) - Y_i) - \tag{26}$$

$$- \sum_1^n c_{ni}^* d_{ni}(t, b), \quad -\infty < b < \infty, \quad 0 < t < 1,$$

$$L_n^*(t\,;\,b) = (n+1)^{-1/2} \sum_{i=1}^n u(t + d_{ni}(t\,;\,b) - Y_i), \quad -\infty < b < \infty, \quad 0 \leqslant t \leqslant 1. \tag{27}$$

Avem :

$$\zeta_n(t\,;\,b) = \sum_{i=1}^n c_{ni}^* d_{ni}(t\,;\,b) = bf(F^{-1}(t)) + b^2\mathbf{0}(n^{-1/2})\,; \tag{28}$$

$$G_n^*(t\,;\,b) = S_n^*(F^{-1}(t)\,;\,b) - \zeta_n(t\,;\,b) = \tag{29}$$

$$= S_n^*(F^{-1}(t)\,;\,b) - bf(F^{-1}(t)) + b^2\mathbf{0}(n^{-1/2}),$$

$$L_n^*(t, b) = (n+1)^{1/2} H_n(F^{-1}(t)\,;\,b), \quad 0 \leqslant t \leqslant 1, \quad -\infty < b < \infty. \tag{30}$$

Teorema 1. (Ghosh, Sen (1972)). *Pentru fiecare $h(>0)$ există două constante pozitive K_1, K_2 și n^* (toate putînd depinde de h) astfel încît pentru $n \geqslant n^*$, $k \geqslant 1$ și $0 < \delta < 1/4$,*

$$P\{\sup_{0<t<1} \sup_{b \in I_n^*} |G_n^*(t, b) - G_n^*(t, 0)| \geqslant K_1 n^{-\delta}(\log n)^k\} \leqslant K_2 n^{-h}, \tag{31}$$

$$P\{\sup_{0<t<1} \sup_{b \in I_n^*} |L_n^*(t\,;\,b) - L_n^*(t, 0)| \geqslant K_1 n^{-\delta}(\log n)^k\} \leqslant K_2 n^{-h}, \tag{32}$$

unde

$$I_n^* = \{b : |b| \leqslant C(\log n)^k, \ 0 < C < \infty\}.$$

Folosim teorema 1 la

Scanned with OKEN Scanner

TEOREMA 2 (Ghosh, Sen (1972)) *In presupunerile* § 1.2, *pentru fiecare* $s(>0)$, *există constantele pozitive* $(c_s^{(1)}, c_s^{(2)})$ *și volumul de selecție* n_s *astfel încît pentru* $\beta = 0$ *și toți* $n \geq n_s$

$$P\{\sup_{b \in I_n^s} |W(\mathbf{X}_n, b)| > c_s^{(1)} n^{-s} (\log n)^{k+1}\} < c_s^{(2)} n^{-s}.$$

Deci, cînd $n \to \infty$

$$\sup_{|b|<c} |W(\mathbf{X}_n, b)| \underset{a.s.}{\to} 0$$

Fie $\mathcal{F}_n$ σ-algebra generată de $\mathbf{R}_n = (R_{n1}, \ldots, R_{nn})$.

De asemenea, fie $T_n^* = C_n T_n$ și presupunem că are loc $H_0 : \beta = 0$. Atunci are loc teorema 3; rezultatul este folosit în demonstrarea *continuității uniforme în probabilitate* a lui $T(X_n)$ în raport cu C_n^{-1}.

TEOREMA 3 (Ghosh, Sen (1972)). *Dacă* $J_n(u) = M(\varphi(U_{ni}))$, $(i-1)/n < u \leq i/n$, $1 \leq i \leq n$, *atunci sub* $H_0 : \beta = 0$, $\{T_n, \mathcal{F}_n\}$ *formează un șir de martingale.*

Observație. Această proprietate a martingalei nu are loc în mod necesar cînd $J_n(i/(n+1)) = \varphi(i/(n+1))$, $1 \leq i \leq n$ [în afară de cazul $\varphi(u) = u$: $0 < u < 1$). De asemenea, cînd $\beta \neq 0$, astfel că date rangurile variabilelor $X_1, \ldots, X_n, X_{n+1}$ nu are o probabilitate constantă $(1/(n+1))$ de a avea rangul i $(1 \leq i \leq n+1)$, proprietatea martingalei anterioare nu are loc în general.

1.4. Proprietăți asimptotice ale intervalului $I_{N(d)}$

Au loc următoarele leme (Ghosh, Sen (1972))

LEMA 1. *Pentru* $(\forall)s(>0)$, $(\exists)$ *constantele* $K_s^{(1)}$, $K_s^{(2)}$ *și un volum de selecție* n_s, *astfel încît pentru* $n \geq n_s$,

$$P\{[C_n(\hat{\beta}_{U,n} - \beta) - A\tau_{\alpha/2}/B(F)] > K_s^{(1)}(\log n)^2\} < K_s^{(2)} n^{-s} \tag{33}$$

$$P\{[C_n(\hat{\beta}_{L,n} - \beta) + A\tau_{\alpha/2}/B(F)] < -K_s^{(1)}(\log n)^2\} < K_s^{(2)} n^{-s} \tag{34}$$

LEMA 2. *Pentru fiecare* $s(>0)$ *există constantele pozitive* $(K_s^{(1)}, K_s^{(2)})$ *și un volum de selecție* n_s, *astfel că pentru* $n \geq n_s$,

$$P\{|C_n(\hat{\beta}_{U,n} - \hat{\beta}_{L,n}) - 2A\tau_{\alpha/2}/B(F)| \geq K_s^{(1)} n^{-s} (\log n)^2\} \leq K_s^{(2)} n^{-s}. \tag{35}$$

LEMA 3. *Pentru fiecare* x $(-\infty < x < \infty)$ *real*

$$\lim_{n \to \infty} P\{C_n(\hat{\beta}_{U,n} - \beta) B(F)/A - \tau_{\alpha/2} \leq x\} = \Phi(x). \tag{36}$$

În final pentru „continuitatea uniformă în probabilitate" a lui $\{\hat{\beta}_{U,n}\}$ în raport cu $\{C_n^{-1}\}$, avem

LEMA 4. *Pentru fiecare* ε *și* η *pozitiv, există un* $\delta(>0)$ *astfel încît pentru* $n \to \infty$

$$P\{\sup_{|n'-n|<\delta n} |C_n(\hat{\beta}_{U,n'} - \hat{\beta}_{U,n})| > \eta\} < \varepsilon \tag{37}$$

și o presupunere similară are loc pentru $\{\hat{\beta}_{L,n}\}$.

Scanned with OKEN Scanner

Proprietăţile asimptotice ale intervalului $I_{N(d)}$ date în următoarea teoremă sînt asemănătoare cu acelea date de Chow, Robbins (1965) şi prezentate în teoremele XXVII.23 şi XXVII.24.

TEOREMA 4 (Ghosh, Sen (1972)). *In presupunerile* § 1.2, $N(d)$ *este o funcţie nedescrescătoare de* $d(>0)$, *este finită,* $MN(d)<\infty$ *pentru toţi* $d>0$, $\lim_{d\to 0} N(d) = \infty$ a.s. şi $\lim_{d\to 0} MN(d) = \infty$. *De asemenea*

$$\lim_{d\to 0} N(d)/Q^{-1}(\nu(d)) = 1 \text{ a.s.}, \tag{38}$$

$$\lim_{d\to 0} P\{\beta \in I_{N(d)}\} = 1-\alpha \text{ pentru toţi } F \in \mathcal{F}(\Psi) \tag{39}$$

$$\lim_{d\to 0} [MN(d)]/Q^{-1}(\nu(d)) = 1, \tag{40}$$

unde

$$\nu(d) = [\{A\tau_{\alpha/2}\}/\{dB(F)\}]^2, \tag{41}$$

§ 2. Teste secvenţiale de rang pentru regresie

2.1. Introducere

În capitolul XXVI § 10 urmînd Sen, Ghosh, (1974) am dezvoltat teste secvenţiale de rang pentru parametrul de locaţie în problema clasică de locaţie în cazul unei selecţii. Aceste rezultate se aplică de asemenea la problema de locaţie în cazul a două selecţii cînd observaţiile sînt luate în pereche. Ghosh, Sen (1977) dezvoltă teste secvenţiale de rang (T.S.R) pentru coeficientul regresiei în modelul regresiei liniare simple. Arată că aceste T.S.R. se termină cu probabilitatea unu în condiţii de regularitate destul de generale.

2.2. T.S.R. propuse

Considerăm modelul de regresie liniară dat de (1) şi presupunem că F are densitatea de repartiţie $f(x)$.

Ne propunem să verificăm ipoteza

(42) $H_0 : \boldsymbol{\beta} = \mathbf{0}$ faţă de $H_1 : \boldsymbol{\beta} = \Delta(>0)$, unde Δ este cunoscut, $\boldsymbol{\beta}$ fiind (β_0, β_1).

Observaţie. Putem avea

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^\circ \text{ faţă de } H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}^\circ + \Delta, \text{ pentru } \boldsymbol{\beta}^\circ \text{ cunoscut.}$$

Totuşi, lucrînd cu $X_i^* = X_i - \beta^\circ c_i$, $i \geq 1$, putem să ajungem întotdeauna la (42).

Statistica de rang folosită bazată pe o selecţie de volum $n(\geq 1)$ este dată de (5).

Scanned with OKEN Scanner

Mai departe, înlocuind în (6) $\mathbf{X}_n = (X_1, \ldots, X_n)$ prin $\mathbf{X}_n - b\mathbf{c}_n = (X_1 - bc_1, \ldots, X_n - bc_n)$ (adică folosind rangul $R_{ni}(b)$ al lui $X_i - bc_i$ printre $X_1 - bc_1, \ldots, X_n - bc_n$) definim statistica de rang corespunzătoare ca

$$(43) \quad T_n(b) = T(\mathbf{X}_n - b\mathbf{c}_n) = \sum_{i=1}^{n} c_{ni}^* J_n((n+1)^{-1} R_{ni}(b)), \quad -\infty < b < \infty.$$

Notăm că după XXXVII.7, $T_n(b)$ este $\downarrow$ în b. Estimatorul $\hat{\beta}_n$ pentru β bazat pe $T_n(b)$ este (Adichie (1967), Ghosh, Sen (1977))

$$(44) \quad \hat{\beta}_n = \frac{1}{2}(\hat{\beta}_{n,1} + \hat{\beta}_{n,2}), \quad n \geq 1$$

unde

$$\hat{\beta}_{n,1} = \sup\{b : T_n(b) > 0\} \text{ și } \hat{\beta}_{n,2} = \inf\{b : T_n(b) < 0\}.$$

Presupunem acum că β_n^* este estimatorul de verosimilitate maximă pentru β cînd F este specificată.
Atunci testul secvențial al raportului de verosimilităţi (Bartlett-Cox) se reduce la pasul n la

$$(45) \quad \Delta C_n^2 I(f)\left[\beta_n^* - \frac{1}{2}\Delta\right].$$

Dacă considerăm şirul

$$(46) \quad \Delta C_n^2 D^2(F)\left[\hat{\beta}_n - \frac{1}{2}\Delta\right] \Big/ v^2, \quad n \geq 1$$

unde $v[D(F)]$ este dată de (XXXVI.121) [18] (am schimbat notaţia, A şi B fiind folosite în descrierea testului secvenţial), atunci din XXXVII. § 4, este natural să înlocuim (45) [(46)] prin următorul şir

$$(47) \quad \Delta C_n D(F)\, T_n\left(\frac{1}{2}\Delta\right) \Big/ v^2, \quad n \geq 1$$

unde din definiţia lui $\hat{\beta}_n$ din (44) şi din teorema 2, înlocuim (în (46) $C_n D(F)\left[\hat{\beta}_n - \frac{1}{2}\Delta\right]$ prin $T_n(\Delta/2)$. Deoarece funcţia de repartiţie F, şi deci, funcţionala $D(F)$ este necunoscută, procedăm ca în § 1. pentru a obţine un estimator puternic consistent D_n pentru $D(F)$. Pentru aceasta definim:

$$(48) \quad \hat{\beta}_{L,n} = \sup\{b : T_n(b) > t_n^*\} \text{ şi } \hat{\beta}_{U,n} = \inf\{b : T_n(b) < -t_n^*\}$$

unde t_n^* satisface
$P_0\{|T_n(0)| \leq t_n^*\} = 1 - \alpha_n$ (α_n specificată şi $\to \alpha$, $0 < \alpha < 1$ cînd $n \to \infty$).
Atunci după lema 2, avem

$$(41) \quad D_n = 2t_n^*/(\hat{\beta}_{U,n} - \hat{\beta}_{L,n}) \xrightarrow{a.s.} D(F) \text{ cînd } n \to \infty.$$

Scanned with OKEN Scanner

Acum T.S.R. propus de Ghosh, Sen (1977) poate fi formulat după cum urmează :

Pentru erorile de tipul I și II $\alpha_1,\ \alpha_2\ \left(0<\alpha_1,\ \alpha_2<\frac{1}{2}\right)$ prestabilite, se consideră două numere

$$A(\leqslant(1-\alpha_2)/\alpha_1) \quad \text{și} \quad B(\geqslant\alpha_2/(1-\alpha_1))$$

și

$$0<B<1<A<\infty.$$

Atunci, se pornește cu o selecție inițială de volum $n_0(\Delta)$ și se definește o variabilă de stopare $N(=N(\Delta))$ ca cel mai mic întreg pozitiv $n(\geqslant n_0(\Delta))$ pentru care următoarea inegalitate

$$b\nu^2<\Delta C_n D_n T_n\left(\frac{1}{2}\Delta\right)<a\nu^2 (b=\log B,\ a=\log A). \tag{50}$$

este stricată.

Dacă $N(\Delta)=n$ și $\Delta C_n D_n T_n\left(\frac{1}{2}\Delta\right)$ este $\leqslant b\nu^2$ (sau $a\nu^2$), acceptăm H_0(sau H_1). Dacă procesul nu se termină, luăm $N=\infty$.

2.3. Terminarea T.S.R. cu probabilitatea unu

Presupunem că (51)

i) $\varphi(u)$ este strict $\nearrow$ $u\in(0,1)$, $\int_0^1 |\varphi(u)|^{2+\delta}\,du<\infty$ pentru $\delta>0$

ii) $\int_0^1 [J_n(u)-\varphi(u)]^2\,du\to 0$ cînd $n\to\infty$ (52)

iii) este satisfăcută (9).

iv) $\liminf_{n\to\infty} n^{-1}C_n^2\geqslant C_0^2,\ 0<C_0^2<\infty$ (53)

v) $\lim_{h\to 0}\sup_{|h_i|\leqslant h, i=1,2}\left\{\int_{-\infty}^{\infty}\varphi'(F(x+h_1))f(x+h_2)\,dF(x)\right\}=D(F)>0$ (54)

unde $D(F)$ este definită în (18).

Are loc

TEOREMA 5 (Ghosh, Sen (1977)) *Sub* (51)—(54) *și* (9), *pentru fiecare* β *fixat*, T.S.R. *se termină cu probabilitatea unu, adică*

$$P_\beta\{N(\Delta)>n\}\to 0 \quad \textit{cînd} \quad n\to\infty. \tag{55}$$

2.4. Funcția C.O. a T.S.R.

Vom presupune că $\Delta \to 0$. De asemenea pentru valori mici ale lui Δ, vom presupune că

$$\beta \in I(\Delta) = \{\beta = \phi\Delta : |\phi| \leqslant K \text{ pentru unii } K > 1\}. \tag{56}$$

De notat că pentru fiecare $\beta > 0$ (sau < 0), funcția C.O. se apropie de o (sau 1) cînd $\Delta \to 0$. Mai departe, vom presupune că volumul inițial al selecției $n_0 = n_0(\Delta)$ satisface

$$\lim_{\Delta \to 0} \Delta^q C^2_{n_0(\Delta)} = \infty, \quad 0 < q < 1, \quad \lim_{\Delta \to 0} \Delta^2 C^2_{n_0(\Delta)} = 0 \tag{57}$$

Fie $\Pi_{J,F}(\phi, \Delta)$ funcția C.O. a T.S.R. bazat pe $N(\Delta)$ din § 2.2. cînd $\beta = \phi\Delta$, φ este funcția scor și F este funcția de repartiție sub cercetare. Cu φ definit ca în §. 1.2. are loc

TEOREMA 6 (Ghosh, Sen (1977)). *Sub* (52)—(54), (9), (56), (57) *și în presupunerea că*

$$\limsup_{n \to \infty} n^{-h} C^2_n \leqslant C^* < \infty \text{ pentru unii } h(\geqslant 1), \tag{58}$$

avem

$$\lim_{\Delta \to 0} \Pi_{J,F}(\phi, \Delta) = P(\phi) = \begin{cases} (A^{1-2\phi} - 1)/(A^{1-2\phi} - B^{1-2\phi}), & \phi \neq \frac{1}{2} \\ a/(a-b), & \phi \neq \frac{1}{2}. \end{cases} \tag{59}$$

Observație. (59) arată că T.S.R. are asimptotic aceeași putere ca și T.S.R.P. (teorema XXVI.22). De asemenea aceasta este adevărată pentru toți φ ce satisfac condițiile din § 1.2. și funcții F absolut continui ce satisfac (54). În particular, pentru $\phi = 0, 1$, (59) se reduce la $1 - \alpha_1, \alpha_2$ astfel că T.S.R. este „asimptotic consistent".

2.5. V.M.V.S. a T.S,R.

Deosebit de situația unei sau a două selecții (XXVI. § 10), aici performanța lui T_n depinde de n împreună cu C^2_n, unde după (53), $n^{-1}C^2_n$ este mărginit (pentru valori mari ale lui n) de o constantă pozitivă, însă $n^{-1}C^2_n$ nu trebuie să tindă la o limită cînd $n \to \infty$. Din acest motiv, este posibil să găsim o valoare asimptotică pentru $\Delta^2 M_\phi C^2_{N(\Delta)}$ cînd $\Delta \to 0$, în condițiile § 2.4, însă trebuie condiții suplimentare pentru a trage concluzia asupra lui $\Delta^2 M_\phi N(\Delta)$.

TEOREMA 7 (Ghosh, Sen (1977)). *Sub* (52)—(54), (9), (56)—(57) *și* (58) *pentru fiecare* $\phi \in I(A)$,

$$\lim_{\Delta \to 0} \{\Delta^2 M_\phi [C^2_{N(\Delta)}]\} = \Psi(\phi, \tau) \tag{60}$$

Scanned with OKEN Scanner

unde M_ϕ *este media calculată cu* $\beta = \phi\Delta$,

$$(61)\quad \Psi(\phi, \tau) = \begin{cases} [a\{1 - P(\phi)\} + bP(\phi)]\left[\tau^2 \Big/ \left(\phi - \frac{1}{2}\right)\right], & \phi \neq \frac{1}{2} \\ \tau^2(-ab), & \phi = \frac{1}{2} \end{cases}$$

şi $\tau^2 = \nu^2/D^2(F)$, $P(\phi)$ *este definit de* (59).

Observaţie. În multe situaţii practice, $\{c_i\}$ aparţin unui interval mărginit şi se poate presupune că

$$(62)\quad n^{-1}C_n^2 \to C^2; \quad 0 < C < \infty \text{ cînd } n \to \infty.$$

Sub (62) şi condiţiile teoremei 7 urmează că

$$(63)\quad \lim_{\Delta \to 0} \Delta^2 M_\phi\{N(\Delta)\} = C^{-2}\Psi(\phi, \tau).$$

Ca în cazul constantelor de regresie egal depărtate [$c_i = c_0 + hi$, $i \geq 1$, $h > 0$], în alte cazuri (62) poate să nu aibă loc; presupunem că

$$(64)\quad C_n^2 = Q(n) \text{ unde } Q(n) \text{ este o funcţie convexă.}$$

Astfel $MQ(N) \geq Q(MN)$ şi deci

$$(65)\quad \lim_{\Delta \to 0} \Delta^2 Q(M_\phi N(\Delta)) \leq \Psi(\phi, \tau)$$

semnul de egal în (65) fiind dificil de atins, deoarece [$\Delta^2 N(\Delta)$] nu este o variabilă degenerată.

2.6. Rezultate privind eficienţa asimptotică relativă

Estimatorul de C.M.M.P. (sau de verosimilitate maximă în teoria normală) pentru β este

$$(66)\quad \beta_n^* = C_n^{-2} \sum_{i=1}^{n} (c_i - \bar{c}_n) X_i, \quad n \geq 1.$$

Dacă punem

$$s_n^2 = \frac{1}{n-2} \sum_{i=1}^{n} (X_i - \bar{X}_n - \beta_n^*(c_i - \bar{c}_n))^2, \quad n > 2$$

atunci dacă înlocuim în (45) $I(f)$ prin s_n^2, procedeul corespunzător este legat de teoria normală a testului secvenţial al raportului de verosimilităţi (XXVI.§ 10.). Acesta se termină cu probabilitatea unu şi cînd $\Delta \to 0$

Scanned with OKEN Scanner

are aceeași funcție C.O. ca cea dată de teorema 6. Pentru acest procedeu, avem paralel cu teorema 7

$$\lim_{\Delta \to 0} \Delta^2 M_\phi C^2_{N(\Delta)} = \Psi(\phi, \tau) \tag{67}$$

unde σ^2 este dispersia adevărată a lui X_i. După (60), (61) și (67), sub (62), eficiența T.S.R. relativă la testul secvențial al raportului de verosimilități, judecată după V.M.V.S. este dată de

$$e = \Psi(\phi, \sigma)/\Psi\phi, r) = \sigma^2/\tau^2 = D^2(F)\sigma^2/\nu^2. \tag{68}$$

Similar, dacă forma actuală a funcției de repartiție F este cunoscută și dacă am folosi T.S.R.P., sub (62), eficientă T.S.R. relativă la T.S.R.P. este

$$e = \Psi(\phi, I^{-1}(f))/\Psi(\phi, \tau) = D^2(F)/(\nu^2 I(f)) \tag{69}$$

(68) și (69) au fost discutate în XXVI. § 10.

Observație. Folosind scorurile normale, adică

$$\varphi(u) = \Phi^{-1}(u),$$

(68) este întotdeauna ≥ 1, unde semnul egal are loc numai cînd F este normală.

§ 3. Teste neparametrice pentru regresia multiplă în censorizarea progresivă

3.1. Introducere

În studiile legate de experiențe clinice și probleme de determinare a duratei de funcționare, cercetătorul întocmește în prealabil planul experimentului ce conține numărul maxim de răspunsuri ce trebuie observate sau durata maximă a experimentului. Deși experimentul este observat continuu în timp, chiar în acest plan cu restricții, cercetătorul are opțiunea să revadă planul după cum progresează experimentul, fiind capabil să termine experimentul la un pas intermediar dacă evidența cumulată indicată de date la acest pas este destul de puternică pentru a respinge ipoteza nulă și dacă continuarea experimentului nu poate conduce la inferență deosebită. Acest procedeu pseudosecvențial (care este desosebit de testul secvențial clasic) este numit schema de censorizare progresivă (punctele de censorizare succesive fiind răspunsurile sau defecțiunile), statisticile test fiind bazate numai pe subiecții necensorizați.

3.2. Noțiuni preliminarii

Fie $\{X_i, i \geq 1\}$ un șir de variabile aleatoare independente cu funcțiile de repartiție $\{F_i, i \geq 1\}$, continui, specificate de modelul :

$$F_i(x) = F(x - \beta_0 - \boldsymbol{\beta}' \mathbf{c}_i), \quad -\infty < x < \infty, \ i \geq 1 \tag{70}$$

unde în acest model convenţional de regresie multiplă, funcţia de repartiţie F nu este cunoscută

$\beta_0, \boldsymbol{\beta}' = (\beta_1, \ldots, \beta_p)$ sînt parametrii necunoscuţi, $p \geqslant 1$ şi $\{\mathbf{c}_i' = (c_{1i}, \ldots, c_{pi}),\ i \geqslant 1\}$ este un şir cunoscut de vectori de regresie.

Ne propunem să verificăm ipoteza :

$$H_0 : \boldsymbol{\beta} = \mathbf{0} \text{ faţă de } H_1 : \boldsymbol{\beta} \neq \mathbf{0} \tag{71}$$

(tratînd β_0 ca un parametru nedorit şi fără a presupune că forma lui F este specificată).

Pentru fiecare $N(\geqslant 1)$, fie $Z_{N1} \leqslant Z_{N2} \leqslant \ldots \leqslant Z_{NN}$ statisticile de ordine corespunzătoare la $(X_1, \ldots, X_N) = \mathbf{X}^{(N)}$; în baza continuităţii funcţiilor F_i, în probabilitate, pot fi neglijate observaţiile cu valori egale. Fie $\mathbf{R}_N = (R_{N1}, \ldots, R_{NN})'$ [$\mathbf{S}_N = (S_{N1}, \ldots, S_{NN})'$] vectorul rangurilor [antiranguri lor] elementelor lui $\mathbf{X}^{(N)}$ astfel că (neglijînd datele cu valori egale)

$$R_{NS_{Ni}} = S_{NR_{Ni}} = i, \quad 1 \leqslant i \leqslant N, \quad X_i = Z_{NR_{Ni}}, \quad 1 \leqslant i \leqslant N.$$

Într-o problemă de determinare a duratei vieţii, se observă succesiv statisticile de ordine $\{Z_{Ni}\}$ în acelaşi timp cu $\{S_{Ni}\}$ corespunzători şi pe baza unei părţi a şirului $\{Z_{Ni}, S_{Ni};\ 1 \leqslant i \leqslant N\}$, verificăm ipoteza H_0 din (71). Testul schemei de censorizare progresivă este un procedeu pseudosecvenţial unde terminarea procedeului este convenabilă dacă se observă $\{Z_{Ni}, S_{Ni}, 1 \leqslant i \leqslant k\}$ pentru $k(\leqslant N)$ şi evidenţa statistică acumulată conduce la o concluzie decisivă. Majumdar, Sen (1978) propune un test în schema de censorizare progresivă bazat pe o alegere convenabilă a statisticilor liniare de rang.

Pentru fiecare $N(\geqslant 1)$ considerăm mulţimea scorurilor

$$\{a_N(1), \ldots, a_N(N)\}$$

generate de funcţia scor

$$\varphi = \{\varphi(u),\ u \in (0, 1)\}$$

în următorul mod :

$$a_N([Nu] + 1) = \varphi_N(u) : 0 < u < 1. \tag{72}$$

unde $[s]$ reprezintă cel mai mare întreg $\leqslant s$ şi

$$\lim_{N \to \infty} \int_0^1 \{\varphi_N(u) - \varphi(u)\}^2 \, du = 0. \tag{73}$$

Mai departe presupunem că $\varphi(u)$ este exprimabilă ca diferenţa a două funcţii $\varphi_1(u)$ şi $\varphi_2(u)$ nedescrescătoare şi de patrat integrabile pe $I = [0, 1]$. În particular, (73) are loc cînd

$$a_N(i) = M\varphi(U_{Ni}) \text{ sau } \varphi(i/(N+1)),\ 1 \leqslant i \leqslant N. \tag{74}$$

Pentru simplificarea notaţiilor fie

$$(75)\qquad \bar{\varphi} = \int_0^1 \varphi(u)\,du = 0,\quad A_\varphi^2 = \int_0^1 \varphi^2(u)\,du - \bar{\varphi}^2 = \int_0^1 \varphi^2(u)\,du = 1,$$

$$\bar{a}_N = \frac{1}{N}\sum_{i=1}^{N} a_N(i) = 0 \text{ şi}$$

$$(76)\qquad A_N^2 = (N-1)\sum_{i=1}^{N} [a_N(i) - \bar{a}_N]^2 = 1, \qquad N \geqslant 2.$$

Acum la pasul k, variabilele aleatoare observate sînt $\{Z_{Ni}, S_{Ni};\ 1 \leqslant i \leqslant k\}$ şi pe baza acestora, definim

$$(77)\qquad \mathbf{T}_{N,k} = (T_{N,k}^{(1)}, \ldots, T_{N,k}^{(p)})' \qquad (1 \leqslant k \leqslant N)$$

luînd

$$\mathbf{T}_{N,k} = \sum_{i=1}^{k} (\mathbf{e}_{S_{Ni}} - \bar{\mathbf{e}}_N)\, a_N(i) + \left[\sum_{i=k+1}^{N} (\mathbf{e}_{S_{Ni}} - \bar{\mathbf{e}}_N)\right]\left[(N-k)^{-1}\sum_{i=k+1}^{N} a_N(i)\right] =$$

$$(78)\qquad = \sum_{i=1}^{k} (\mathbf{e}_{S_{Ni}} - \bar{\mathbf{e}}_N)\,[a_N(i) - a_N^*(k)]$$

unde $\bar{\mathbf{e}}_N = N^{-1}\sum_{i=1}^{N} \mathbf{e}_i$ şi

$$(79)\qquad a_N^*(k) = \begin{cases} (N-k)^{-1}\sum_{i=k+1}^{N} a_N(i), & 1 \leqslant k \leqslant N-1 \\ 0 \qquad\qquad , & k = 0, N \end{cases}$$

şi convenţional, fie $\mathbf{T}_{N,0} = \mathbf{0}$. Notăm că

$$\mathbf{T}_{N,N-1} = \mathbf{T}_{NN} = \mathbf{T}_N,$$

unde

$$(80)\qquad \mathbf{T}_N = \sum_{i=1}^{N} (\mathbf{e}_{S_{Ni}} - \bar{\mathbf{e}}_N)\, a_N(i) = \sum_{i=1}^{N} (\mathbf{e}_i - \bar{\mathbf{e}}_N)\, a_N(R_{Ni}).$$

Dacă luăm

$$(81)\qquad b_{N,k}(i) = \begin{cases} a_N(i), & 1 \leqslant i \leqslant k \\ a_N^*(k), & k < i \leqslant N \end{cases} \qquad k = 0, 1, \ldots, N$$

atunci putem să scriem $\mathbf{T}_{N,k}$ ca

$$(83)\qquad \mathbf{T}_{N,k} = \sum_{i=1}^{N} (\mathbf{e}_i - \bar{\mathbf{e}}_N)\, b_{N,k}(R_{Ni}), \qquad k = 0, 1, \ldots, N$$

Scanned with OKEN Scanner

Notăm că după (76), (79) și (81) pentru ficcare $1 \leqslant k \leqslant q \leqslant N$,

$$(N-1)^{-1} \sum_{i=1}^{N} b_{N,k}(i)\, b_{N,q}(i) =$$

$$= (N-1)^{-1} \left\{ \sum_{i=1}^{k} a_N^2(i) + (N-k)\,[a_N^*(k)]^2 \right\} =$$

(83)

$$= A_N^2 - (N-1)^{-1} \sum_{i=k+1}^{N} [a_N(i) - a_N^*(k)]^2 =$$

$$= 1 - (N-1)^{-1} \sum_{i=k+1}^{N} [a_N(i) - a_N^*(k)]^2 = A_{N,k}^2 \ (notație)$$

unde $A_{N,k}^2$ este $\nearrow$ în k $(0 \leqslant k \leqslant N)$ și

$$A_{N,0}^2 = 0, \quad A_{N,N-1}^2 = A_{N,N}^2 = A_N^2 = 1.$$

De asemenea notăm

(84) $$\mathbf{C}_N = \sum_{i=1}^{N} (\mathbf{e}_i - \bar{\mathbf{e}}_N)(\mathbf{e}_i - \bar{\mathbf{e}}_N)' = ((c_{Njl}))_{j,l,\ldots,p}.$$

și presupunem că există in număr pozitiv N_0 astfel incît

(85) $\mathbf{C}_N$ este pozitiv definită $(\forall) N \geqslant N_0$.

Sub H_0 din (71),

$$\mathbf{Z}_N = (Z_{N1}, \ldots, Z_{NN})' \text{ și } \mathbf{R}_N \text{ (sau } \mathbf{S}_N)$$

sînt stochastic independente; $\mathbf{R}_N$ (sau $\mathbf{S}_N$) iau toate permutările posibile ale lui $(1, \ldots, N)$ cu probabilitate egală $(N!)^{-1}$. Deci, din (82) la (84) obținem că

(86) $$M(\mathbf{T}_{N,k} \mid H_0) = 0, \qquad 0 \leqslant k \leqslant N$$

(87) $$M(\mathbf{T}_{N,k}\mathbf{T}_{N,q}' \mid H_0) = A_{N,k \wedge q}^2 \cdot C_N, \qquad 0 \leqslant k, q \leqslant N$$

unde $a \wedge b = \min(a, b)$. Definim

(88) $$L_{N,k} = \begin{cases} 0, & k = 0 \\ (\mathbf{T}_{N,k}' \mathbf{C}_N^{-1} \mathbf{T}_{N,k})^{1/2}, & k = 1, \ldots, N. \end{cases}$$

Testele în schemele de censorizare progresivă (S.C.P.) sînt bazate pe șirul parțial:

(89) $$\{L_{N,k} : 0 \leqslant k \leqslant r\}$$

Scanned with OKEN Scanner

unde r este orice întreg prestabilit, astfel încît

$$r/N \to \delta : 0 < \delta \leqslant 1 \text{ cînd } N \to \infty \tag{90}$$

În realitate δ este obișnuit mai mic decît 1.

3.3. Teste în S.C.P.

Majundar, Sen (1978) consideră următoarele statistici test :
i) statistici de tip Kolmogorov-Smirnov (KS)

$$K^*_{N,r} = \max_{0 \leqslant k \leqslant r} \{A^{-1}_{N,r} L_{N,k}\} = A^{-1}_{N,r} \{\max_{0 \leqslant k \leqslant r} (\mathbf{T}'_{N,k} \mathbf{C}_N^{-1} \mathbf{T}_{N,k})^{1/2}\} \tag{91}$$

ii) statistici de tip Cramer-von Mises

$$M^*_{N,r} = \sum_{k=0}^{r-1} \lambda^*_k L^2_{N,k} / A^2_{N,r} = A^{-2}_{N,r} \sum_{k=0}^{r-1} \lambda^*_k (\mathbf{T}'_{N,k} \mathbf{C}_N^{-1} \mathbf{T}_{N,k}), \tag{92}$$

unde

$$\lambda^*_k = A^{-2}_{N,r}[(N-k-1)/(N-1)(N-k)]\,[a_N(k+1) - a^*_N(k+1)]^2 \tag{93}$$

pentru $k = 0, 1, \ldots, r-1$.

Aceste două statistici pot fi exprimate în termeni de procese stocastice după cum urmează.

Pentru fiecare $N \geqslant r \geqslant 1$ introducem șirul de funcții $\{k_{N,r}(t),\ t \in I\}$ cu valori întregi, nedescrescător și continui la dreapta, unde

$$k_{N,r}(t) = \max \{k : A^2_{N,k} \leqslant t A^2_{N,r}\}, \qquad t \in I \tag{94}$$

și $A^2_{N,k}$ este definită de (83). Considerăm atunci procesul $Y_{N,r} = \{Y_{N,r}(t),\ t \in I\}$ scriind

$$Y_{N,r}(t) = A^{-1}_{N,r} L_{N,k_{N,r}}(t), \qquad t \in I \tag{95}$$

Astfel, $Y_{N,r}$ aparțin spațiului $D[0, 1]$, înzestrat cu topologia J_1 a lui Skorokhod. Atunci, avem :

$$K^*_{N,r} = \sup_{t \in I} Y_{N,r}(t) \tag{96}$$

$$M^*_{N,r} = \int_0^1 Y^2_{N,r}(t)\, dt. \tag{97}$$

Dacă luăm

$$K^{(k)}_{N,r} = A^{-1}_{N,r} \{\max_{0 \leqslant q \leqslant k} (\mathbf{T}'_{N,q} \mathbf{C}_N^{-1} \mathbf{T}_{N,q})^{1/2}\}, \qquad 0 \leqslant k \leqslant r \tag{98}$$

Scanned with OKEN Scanner

și

$$(99)\qquad M_{N,r}^{(k)} = A_{N,r}^{-2}\sum_{s=0}^{k-1}\lambda_s^*(\mathbf{T}'_{N,s}\mathbf{C}_N^{-1}\mathbf{T}_{N,s}) = \int_0^{t_{N,r}^{(k)}} Y_{N,r}^2(t)\,dt, \qquad 0\leqslant k\leqslant r$$

unde $t_{N,r}^{(k)} = A_{N,k}^2/A_{N,r}^2$ pentru $k = 0, 1, \ldots, r$, avem

$$(100)\qquad K_{N,r}^{(k)} \nearrow \text{ în } k \text{ și } M_{N,r}^{(k)} \nearrow \text{ în } k, \text{ pentru } 0\leqslant k\leqslant r\ (\leqslant N).$$

Atunci proprietatea de monotonicitate din (100) este încorporată în următoarea formulare a testelor în S.C.P. Pentru un prag de semnificație prestabilit $\alpha(0<\alpha<1)$, fie $K_{N,r,\alpha}^*$ și $M_{N,r,\alpha}^*$ definiți de

$$(101)\qquad P\{K_{N,r}^* \geqslant K_{N,r,\alpha}^* \mid H_0\} \geqslant \alpha > P\{K_{N,r}^* > K_{N,r,\alpha}^* \mid H_0\};$$

$$(102)\qquad P\{M_{N,r}^* \geqslant M_{N,r,\alpha}^* \mid H_0\} \geqslant \alpha > P\{M_{N,r}^* > M_{N,r,\alpha}^* \mid H_0\}.$$

Atunci, pentru N articole supuse unui test de viață, la defecțiunile succesiv observate, în fiecare defecțiune Z_{Nk}, calculăm $K_{N,r}^{(k)}$ (sau $M_{N,r}^{(k)}$), $k\geqslant 0$. Dacă pentru prima dată, pentru un $k(\geqslant r)$, $K_{N,r}^{(k)}$ este $>K_{N,r,\alpha}^*$ (sau $M_{N,r}^{(k)}$ este $>M_{N,r,\alpha}^*$) experimentarea este oprită și respingem H_0. Dacă un astfel de $k(\leqslant r)$ nu există, experimentarea este tăiată cu acceptarea ipotezei H_0. Prin definiție $K_{N,r}^*$ și $M_{N,r}^*$ sînt funcții de $\mathbf{R}_n$(și $\mathbf{C}_1, \ldots, \mathbf{C}_N$) și deci, sub H_0 din (71), sînt statistici cu repartiții libere. După (82), (88), (93) și (98) aceste statistici rămîn invariante sub orice transformare nesingulară asupra vectorilor de regresie $\{\mathbf{c}_i\}$. Dacă $\mathbf{d}_i = \mathbf{c}_0 + \Gamma\mathbf{c}_i$, $i\geqslant 1$, unde $\mathbf{c}_0$ este arbitrar și Γ este nesingulară, și pentru (70) rescriem

$$x - \beta_0 - \boldsymbol{\beta}'\mathbf{c}_i \text{ ca } x - \gamma_0 - \boldsymbol{\gamma}'\mathbf{d}_i, \qquad 1\leqslant i\leqslant N$$

unde

$$\boldsymbol{\gamma}' = \boldsymbol{\beta}'\Gamma^{-1} \text{ și } \gamma_0 = \beta_0 - \boldsymbol{\gamma}'\mathbf{c}_0,$$

atunci înlocuind în (80), (82) și (88), $\mathbf{c}_i$ prin $\mathbf{d}_i$ și notînd statisticile rezultate prin $\tilde{L}_{N,k}$ avem

$$L_{N,k} = \tilde{L}_{N,k}, \ (\forall)\, k\geqslant 0 \text{ și } \Gamma \text{ nesingulară.}$$

Deoarece în multe cazuri (problema de locație în cazul multiselecției), $\mathbf{c}_i$ și $\boldsymbol{\beta}_i$ din (70) nu sînt unic definite, această invariantă este importantă și elimină acest arbitrar în alegerea lui $\mathbf{c}_i$. Astfel, testele în S.C.P. sînt invariante, și statisticile pe care se bazează au repartiții libere. Pentru valori mici ale lui N, repartiția nulă a statisticii $K_{N,r}^*$ sau $M_{N,r}^*$ poate fi determinată prin enumerarea directă a repartiției exacte a statisticii $\mathbf{R}_N$ (sau $\mathbf{S}_N$) pe cele $N!$ permutări egal probabile ale lui $(1, \ldots, N)$, ce dau $N^{[r]}$ realizări egal probabile ale lui $(S_{N1}, \ldots, S_{Nr})$ din $(1, \ldots, N)$.
Calculele devin laborioase cînd r (și deci și N) are valori mari. Deci prezintă importanță repartiția asimptotică.

3.4. Repartiția asimptotică în ipoteza nulă

Pentru a studia repartițiile asimptotice ale statisicilor $K^*_{N,r}$ și $M^*_{N,r}$ (amîndouă fiind funcționale de procesul $Y_{N,r}$ definit în (94), (95)) vom considera mai întîi convergența slabă a $\{Y_{N,r}\}$ către funcționalele unei mișcări browniene. Fie

$$W_j = \{W_j(t),\ t \in I\}, \qquad j = 1, \ldots, p$$

p copii independente ale mișcării Browniene standard pe I și definim

$Y = \{Y(t),\ t \in I\}$, prin

$$Y(t) = \left[\sum_{j=1}^{p} W_j^2(t)\right]^{1/2}, \qquad t \in I. \tag{103}$$

Atunci Y aparține spațiului $C[0, 1]$ cu probabilitatea 1. La acest pas introducem următoarea condiție (de tip Noether)

$$\limsup_{N\to\infty} \{\max_{1\leqslant i\leqslant N} (\mathbf{c}_i - \bar{\mathbf{c}}_N)' \mathbf{C}_N^{-1} (\mathbf{c}_i - \bar{\mathbf{c}}_N)\} = 0. \tag{104}$$

Atunci are loc

TEOREMA 8 (Majundar, Sen (1978)). *Sub H_0 cu* (71), (90), (104) *și presupunerile făcute asupra scorurilor din* § 3.2., *cînd* $N \to \infty$

$$Y_{N,r} \xrightarrow{R} Y \text{ în topologia } J_1 \text{ pe } D[0, 1] \tag{105}$$

Fie $\mathscr{B}_{N,k}$ sigma algebră generată de $\mathbf{S}_N^{(k)} = (\mathbf{S}_{N1}, \ldots, \mathbf{S}_{Nk})$, $0 \leqslant k \leqslant N$, $\mathscr{B}_{N,0}$ fiind sigma algebră trivială. Atunci pentru fiecare $N(\geqslant 1)$, $\mathscr{B}_{N,k}$ este nedescrescătoare în k. Mai departe definim N_0 ca în (79).

LEMA 5. *Sub H_0 din* (71), $\{T_{N,k}, \mathscr{B}_{N,k};\ 0 \leqslant k \leqslant N\}$ *pentru fiecare* $N(\geqslant 1)$ *este un martingal și* $\{L_{N,k}, \mathscr{B}_{N,k};\ 0 \leqslant k \leqslant N\}$ *este un submartingal.*

Pentru fiecare $t : 0 < t < 1$, definim

$$\nu(t) = \int_0^t \varphi^2(u)\, du + (1 - t)^{-1} \left(\int_t^1 \varphi(u)\, du\right)^2, \tag{106}$$

astfel că $\nu(t)$ este $\uparrow$ în $t \in (0, 1)$, $\nu(0) = \lim_{t\downarrow 0} \nu(t) = 0$ și $\nu(1) = \lim_{t\uparrow 1} \nu(t) = A_\varphi^2 = 1$.

LEMA 6. *Sub* (72) — (76), $k/N \to t : t \in [0, 1]$ *asigură că*

$$A_{N,k}^2 = (N - 1)^{-1} \sum_{i=1}^{N} b_{N,k}^2(i) \to \nu(t), \text{ cînd } N \to \infty, \tag{107}$$

unde $b_{N,k}(i)$ *sînt definite în* (81).

Scanned with OKEN Scanner

Acum introducem

$$\omega_p^* = \sup_{t\in I} |\, Y(t)\,| = \sup_{t\in I}\left[\sum_{j=1}^{p} W_j^2(t)\right]^{1/2}, \tag{108}$$

$$\omega_p^0 = \int_0^1 Y^2(t)\,dt = \sum_{j=1}^{p}\int_0^1 W_j^2(t)\,dt = \sum_{j=1}^{p}\omega_j, \tag{109}$$

unde ω_j sînt variabile aleatoare nenegative independente și identic repartizate. Din teorema 8, urmează că

$$M_{N,r}^* \xrightarrow{R} \omega_p^0 \text{ și } K_{N,r}^* \xrightarrow{R} \omega_p^* \text{ cînd } N \to \infty. \tag{110}$$

Funcția caracteristică $g(\theta)$ a lui ω_1 (sau orice ω_j, $j \geqslant 1$) este dată de (Dugue (1969))

$$g(\theta) = (\cos(2i\theta)^{1/2})^{-1/2} = \prod_{k=1}^{\infty}\{1 - 2i\theta u_k\}^{-1/2}, \tag{111}$$

$$u_k = 4/\pi^2(2k-1)^2, \qquad k \geqslant 1.$$

Prin urmare, funcția caracteristică $g_p^0(\theta)$ a lui ω_p^0 este dată de

$$g_p^0(\theta) = [g(\theta)]^p = \prod_{k=1}^{\infty}\{1 - 2i\theta u_k\}^{-p/2}. \tag{112}$$

Dacă $\{U_j, j \geqslant 1\}$ este un șir de variabile aleatoare independente, identic repartizate, unde $U_j \sim \chi^2(p)$, atunci pentru fiecare c,

$$M\{\exp[itcU_j]\} = (1 - 2itc)^{-p/2} \tag{113}$$

astfel că din (112) și (113) avem:

$$\omega_p^0 \overset{R}{=} \sum_{k=1}^{\infty}\{4/\pi^2(2k-1)^2\}\,U_k, \tag{114}$$

unde $\overset{R}{=}$ reprezintă egalitatea repartițiilor. Din (110) și (114), sub H_0, cînd $N \to \infty$

$$M_{N,r}^* \xrightarrow{R} \sum_{k=1}^{\infty}\{4/\pi^2(2k-1)^2\}\,U_k = U_0^0 \ (\textit{notație}) \tag{115}$$

Repartiția lui U_0^0 nu este cunoscută; Majumdar, Sen (1978) obțin percentilele empirice prin simulare (tabelele 1 și 2).

Tabelul 1

Valorile simulate ale repartiției lui ω_p^* în ipoteza nulă pentru valori ale lui p și α

	p			
α	1*	2	3	4
0,01	2,81	3,22	3,71	3,89
0,05	2,24	2,70	3,05	3,31
0,10	1,96	2,35	2,78	3,04

* Exact

Tabelul 2

Valorile simulate ale repartiției lui ω_p^0 în ipoteza nulă

	p			
α	1	2	3	4
0,10	2,87	4,06	5,57	6,54
0,05	1,66	2,67	3,49	4,33
0,10	1,19	2,01	2,80	3,64

Scanned with OKEN Scanner

Pentru $p = 1$, repartiția lui $\omega_1^* = \sup_{t \in I} |W_1(t)|$ este cunoscută și este dată de (Billingsley, p. 79)

$$(116)\quad P\{\omega_1^* \leqslant x\} = \sum_{k=-\infty}^{\infty} (-1)^k [\Phi((2k+1)x) - \Phi((2k-1)x)], \quad x \geqslant 0.$$

Prin definiție $\{Y(t),\ t \in I\}$ este un submartingal, astfel că pentru fiecare $x > 0$,

$$(117)\quad \begin{aligned} P\{\sup_{t \in I} Y(t) > x\} &= P\{\sup_{t \in I} e^{\theta Y^2(t)} > e^{\theta x^2}\}, \qquad \theta > 0 \\ &\leqslant \inf_{\theta > 0} \{e^{-\theta x^2} M(e^{\theta Y^2(t)})\} = \\ &= \inf_{\theta > 0} \{e^{-\theta x^2}(1 - 2\theta)^{-p/2}\} \end{aligned}$$

deoarece $Y^2(1) = \sum_{j=1}^{p} W_j^2(1) \sim \chi^2(p)$; inegalitatea din (117) este bazată pe inegalitatea lui Kolmogorov pentru submartingale. Pentru $x^2 > p$, membrul drept se reduce la

$$(118)\quad (e/p)^{p/2} x^p e^{-x^2/2} (\simeq (2\pi)^{1/2} \{(p/2)^{1/2} 2^{p/2} \sqrt{p/2}\}^{-1} x^p e^{-x^2/2})$$

Pe de altă parte, cum $\omega_p^* > \omega_1^*$, din (116), pentru fiecare $x > 0$, avem:

$$(119)\quad \begin{aligned} P\{\omega_p^* > x\} &\geqslant P\{\omega_1^* > x\} = 1 - P\{\omega_1^* \leqslant x\} = \\ &= 2[1 - \Phi(x)] + 2 \sum_{k=1}^{\infty} (-1)^{k-1} [\Phi((2k+1)x) - \Phi((2k-1)x)] \\ &\simeq 4[1 - \Phi(x)] \text{ cînd } x \text{ este mic.} \end{aligned}$$

Din (117)—(119) obținem că

$$(120)\quad \lim_{x \to \infty} [-2x^{-2} \log P\{\omega_p^* > x\}] = 1 \text{ pentru fiecare } p \geqslant 1.$$

3.5. Repartiția asimptotică în ipotezele nenule

Presupunem că există $\{X_{Ni},\ 1 \leqslant i \leqslant N,\ N \geqslant 1\}$, independente, unde pentru fiecare N,

$$(121)\quad P\{X_{Ni} \leqslant x\} = F(x - \beta_0 - \boldsymbol{\beta}' \mathbf{c}_{Ni}^*), \qquad 1 \leqslant i \leqslant N,\ -\infty < x < \infty$$

unde $\mathbf{c}_{Ni}^* = (c_{N1i}^*, \ldots, c_{Npi}^*)'$, $i = 1, \ldots, N$ și

$$(122)\quad \sum_{i=1}^{N} \mathbf{c}_{Ni}^* = 0 \quad \text{și} \quad \sum_{i=1}^{N} \mathbf{c}_{Ni}^* \mathbf{c}_{Ni}^{*\prime} = 1.$$

Scanned with OKEN Scanner

și în acest caz, (104) se reduce la

$$\max_{1\leqslant j\leqslant p}\{\max_{1\leqslant i\leqslant N} c_{Nji}^{*2}\}\to 0.$$

La ipoteza $H_0:\boldsymbol{\beta}=\mathbf{0}$ considerăm șirul de ipoteze alternative $\{H_N\}$,

(123) $$H_N:(121),\ (122)\ \text{au loc cu}\ \boldsymbol{\beta}\neq\mathbf{0}.$$

Presupunem că F din (121) are o densitate absolut continuă f cu informația Fisher finită.

Definim $\varphi(u,f)$ ca în (XXXVI.119).

De asemenea, pentru $s\in(0,1)$ fie

(124) $$\varphi_s(u)=\begin{cases}\varphi(u), & 0<u\leqslant s\\ (1-s)^{-1}\int_s^1\varphi(t)\,dt=\varphi_s^*, & s<u<1,\end{cases}$$

și definim $\varphi_s(u,f)$, $0<u<1$ în același mod. Pentru $s\in(0,1)$, fie

(125) $$\rho(s)=\left(\int_0^1\varphi(u)\,\varphi(u,f)\,du\right)\Big/I^{1/2}(f)=\left(\int_0^1\varphi_s(u)\,\varphi_s(u,f)\,du\right)\Big/I^{1/2}(f),$$

și definim $\nu(s)$, $0\leqslant(\nu)s\leqslant s$ ca în (106). Atunci are loc

TEOREMA 9 (Majumdar, Sen (1978)). *Sub* (121)—(123) *și condițiile asupra funcției scor din* § 3.2, $\widetilde{\mathbf{W}}_{N,r}$ definit de

$$\widetilde{\mathbf{W}}_{N,r}=\{\mathbf{W}_{N,r}(t),\ t\in I\}=\{(\widetilde{W}_{N,r}^{(1)}(t),\ \ldots,\ \widetilde{W}_{N,r}^{(p)}(t))',\ t\in I\}$$

prin

(126) $$\widetilde{\mathbf{W}}_{N,r}(t)=A_{N,r}^{-1}\widetilde{\mathbf{T}}_{N,k_{N,r}}(t),\ t\in I,$$

unde $\widetilde{\mathbf{T}}_{N,k}$ se obține din (82) prin înlocuirea lui $\mathbf{c}_i-\bar{\mathbf{c}}_N$ prin $\tilde{\mathbf{c}}_i$, $1\leqslant i\leqslant N$, cu

(127) $$\tilde{\mathbf{c}}_i=\mathbf{E}_N(\mathbf{c}_i-\bar{\mathbf{c}}_N),\ 1\leqslant i\leqslant n\ \text{astfel încît}\ \sum_{i=1}^N\tilde{\mathbf{c}}_i=0,\ \sum_{i=1}^N\tilde{\mathbf{c}}_i\tilde{\mathbf{c}}_i'=\mathbf{I}_p,$$

$\mathbf{E}_N$ fiind matricea nesingulară pentru care

(128) $$\mathbf{E}_N'\mathbf{C}_N\mathbf{E}_N=\mathbf{I}_p=\text{Diag}\,(1,\ \ldots,\ 1)$$

converge în repartiție în topologia J_1 pe $D[0,1]$ la o variabilă p-dimensională Gaussiană $\mathbf{W}+\mu$, unde $\mathbf{W}$ este definită de

(129) $$\widetilde{\mathbf{W}}_{N,r}\xrightarrow{R}W\ \text{în topologia}\ J_1\ \text{pe}\ D(0,1)$$

Scanned with OKEN Scanner

și $\boldsymbol{\mu} = \{\boldsymbol{\mu}(t);\ t \in I\}$ este dată de

$$\boldsymbol{\mu}(t) = \boldsymbol{\beta}\rho(\nu^{-1}(t\nu(\delta))\,[I(f)/\nu(\delta)]^{1/2}, \qquad 0 \leqslant t \leqslant 1. \tag{130}$$

Observație. Cum

$$\mathbf{W}(t) = (W_1(t), \ldots, W_p(t))', \quad t \in I$$

$$\boldsymbol{\mu}(t) = (\mu_1(t), \ldots, \mu_p(t))', \qquad t \in I,$$

din (93), (96), (126),

$$Y_{N,r}(t) = ([\widetilde{\mathbf{W}}_{N,r}(t)]'\,[\widetilde{\mathbf{W}}_{N,r}(t)])^{1/2} \tag{131}$$

și teorema 9 obținem că (sub $\{H_N\}$)

$$K^*_{N,r} \xrightarrow{R} \sup_{t \in I} \left\{\left[\sum_{j=1}^{p} (W_j(t) + \mu_j(t))^2\right]^{1/2}\right\}, \tag{132}$$

$$M^*_{N,r} \xrightarrow{R} \sum_{j=1}^{p} \int_0^1 \{W_j(t) + \mu_j(t)\}^2\,dt. \tag{133}$$

$\mu_j(t)$ definite de (130) nu sînt în general funcții liniare de t și ca în § 3.4, este dificil de obținut repartițiile exacte pentru membri din dreapta ai relațiilor (132) și (133). Ca rezultat este dificil de a obține concluzii privind performanțele relative ale acestor teste în S.C.P. pentru alternative continui. Pentru acest motiv, folosim eficiența în sensul lui Bahadur, în scopul comparării diferitelor statistici test și funcții scor.

3.6. Eficiența asimptotică relativă în sensul lui Bahadur pentru teste în S.C.P.

Mai întîi, asemănător cu (120) obținem un rezultat limită pentru coada repartiției statisticii de tip Cramer-von Mises. După (115)

$$\frac{4}{\pi^2}\,U_1 < U_0^0 = \frac{4}{\pi^2}\,U_1 + \sum_{k=2}^{\infty} \{4/\pi^2(2k-1)^2\}\,U_k \tag{134}$$

unde fiecare $U_j \sim \chi^2(p)$, $j \geqslant 1$, astfel încît pentru fiecare $\lambda > 0$,

$$P\{U_j > \lambda^2\} = (2^{p/2}\sqrt{p/2})^{-1} \int_{\lambda^2}^{\infty} e^{-x^2/2}(x^2)^{p/2-1}\,dx^2. \tag{135}$$

Urmează că

$$\lim_{\lambda \to \infty} [-(2/\lambda^2) \log P\{U_j > \lambda^2\}] = 1, \tag{136}$$

astfel că din (134) și (136) obținem că

$$\limsup_{\lambda\to\infty}\,[-(2/\lambda^2)\log P\{U_0^0>\lambda^2\}]\leqslant$$

(137)

$$\leqslant\limsup_{\lambda\to\infty}\,[-(2/\lambda^2)\log P\{U_1>\lambda^2\pi^2/4\}]=\pi^2/4.$$

Pe de altă parte, din independența lui U_j,

$$P\left\{\sum_{i=1}^{\infty}\{4/\pi^2(2k-1)^2\}\,U_k>\lambda^2\right\}=$$

$$=P\left\{\exp\left[\theta\sum_{k=1}^{\infty}\{4/\pi^2(2k-1)^2\}\,U_k\right]>\exp[\theta\lambda^2]\right\}\underset{\theta>0}{\leqslant}$$

$$\leqslant\inf_{\theta>0}\left\{e^{-\theta\lambda^2}M\left[\exp\theta\sum_{k=1}^{\infty}\{4/\pi^2(2k-1)^2\}\,U_k\right]\right\}=$$

(138)

$$=\inf_{\theta>0}\left\{e^{-\theta\lambda^2}\prod_{k=1}^{\infty}M(e^{\theta U_k 4/\pi^2(2k-1)^2})\right\}=$$

$$=\inf_{\theta>0}\left\{e^{-\theta\lambda^2}\prod_{k=1}^{\infty}(1-8\,\theta/\pi^2(2k-1)^2)^{-p/2}\right\}\underset{\varepsilon>0}{\leqslant}$$

$$\leqslant[c(\varepsilon)]\exp\{-\pi^2\lambda^2/8+\varepsilon\lambda^2\}$$

unde $c(\varepsilon)(<\infty)$ depinde de $\varepsilon(0<\varepsilon<\pi^2/8)$. Astfel, alegînd $\varepsilon(>0)$ arbitrar de mic, din (138), obținem că

$$\text{(139)}\quad\liminf_{\lambda\to\infty}\left[-\frac{2}{\lambda^2}\log P\{U_0^0>\lambda^2\}\right]\geqslant\frac{\pi^2}{4}-2\varepsilon,\quad(\forall)\varepsilon(0<\varepsilon<\pi^2/8)$$

astfel că din (137) și (139)

$$\text{(140)}\qquad\lim_{\lambda\to\infty}\,[-(2/\lambda^2)\log P\{U_0^0>\lambda^2\}]=\pi^2/4.$$

În unele cazuri specifice, poate fi studiată convergența stochastică a lui $L_{N,k}$, $k\leqslant N$ pentru alternative generale (Chatterjee, Sen (1973) consideră problema celor două selecții). Însă în general, sînt cerute condiții de regularitate suplimentare. Presupunem că $X_1,\ldots,X_n$ sînt independente cu funcțiile de repartiție $F_1,\ldots,F_N$ și definind $\tilde{e}_i$ ca în (127), luăm

$$\text{(141)}\quad\bar{F}_{(N)}(x)=N^{-1}\sum_{i=1}^{N}F_i(x),\ \mathbf{D}_{(N)}(x)=N^{-1/2}\sum_{i=1}^{N}\tilde{e}_iF_i(x_i),\ -\infty<x<\infty,$$

Scanned with OKEN Scanner

și definind $\mathbf{C}_N$ prin (84), presupunem că

(142) i) $\lim_{N\to\infty} N^{-1}\mathbf{C}_N = \mathbf{C}$ și $\lim_{N\to\infty} \bar{c}_N = \bar{c}$ există,

(143) ii) $\lim_{N\to\infty} \bar{F}_{(N)}(x) = \bar{F}(x)$ există pentru toți x

(144) iii) $\lim_{N\to\infty} \bar{\mathbf{D}}_{(N)}(x) = \mathbf{D}(x)$ există pentru toți x.

De asemenea definim $\nu(t)$ ca în (106) și fie

$$\tau(x) = \int_{-\infty}^{x} \varphi(\bar{F}(y))\,\mathrm{d}\mathbf{D}(y) - \{1 - \bar{F}(x)\}^{-1}\mathbf{D}(x)\left(\int_{x}^{\infty} \varphi(\bar{F}(x))\,\mathrm{d}\bar{F}(x)\right), \tag{145}$$

$$\xi(x) = (\tau(x))'(\tau(x)), \qquad -\infty < x < \infty \tag{146}$$

$$\xi^*(x) = \sup_{-\infty<y\le x} [\xi(y)]^{1/2}, \qquad -\infty < x < \infty \tag{147}$$

$$\xi^0(x) = \left(\int_{-\infty}^{x} \xi(y)\,\mathrm{d}\nu(\bar{F}(y))\right) \Big/ \nu(\bar{F}(x)), \qquad -\infty < x < \infty. \tag{148}$$

Atunci, sub (72)—(75), (90) și (142)—(145), când $N \to \infty$

$$K^*_{N,r} \to \xi^*(\bar{F}^{-1}(\delta))/\nu^{1/2}(\delta), \quad M^*_{N,r} \to \xi^0(\bar{F}^{-1}(\delta)), \tag{149}$$

$$I^2_{N,r}/A^2_{N,r} \to \xi(\bar{F}^{-1}(\delta))/\nu(\delta). \tag{150}$$

În particular, pentru modelul (70), dacă $\mathbf{c}_i$ și toate elementele mărginite și $\boldsymbol{\beta}$ este apropiat de $\mathbf{0}$, atunci (149) și (150) se simplifică la

$$K^*_{N,r} \xrightarrow{\text{a.s.}} (\boldsymbol{\beta}'\mathbf{C}^{-1}\boldsymbol{\beta})^{1/2}\left\{\sup_{0<t\le\delta}\left[\int_0^1 \varphi_t(u)\varphi(u,t)\,\mathrm{d}u\right]^2 \Big/ \nu(\delta)\right\}^{1/2} + o(\|\boldsymbol{\beta}\|) =$$

$$= [(\boldsymbol{\beta}'\mathbf{C}^{-1}\boldsymbol{\beta})I(f)/\nu(\delta)]^{1/2}\{\sup_{0<t\le\delta}\rho^2(t)\}^{1/2} + o(\|\boldsymbol{\beta}\|), \tag{151}$$

unde $\rho(s)$ este definit de (125) și $\|\boldsymbol{\beta}\| = \boldsymbol{\beta}'\boldsymbol{\beta}$;

$$M^*_{N,r} \xrightarrow{\text{a.s.}} (\boldsymbol{\beta}'\mathbf{C}^{-1}\boldsymbol{\beta})\left[\int_{-\infty}^{\bar{F}^{-1}(\delta)} \rho^2(\bar{F}(y)\,\mathrm{d}\nu(\bar{F}(y))\right][I(f)/\nu^2(\delta)] + o(\|\boldsymbol{\beta}\|) =$$

$$= [(\boldsymbol{\beta}'\mathbf{C}^{-1}\boldsymbol{\beta})\,I(f)/\nu(\delta)]\left[\int_0^{\delta} \rho^2(u)\,\mathrm{d}\nu(u)\right]/\nu(\delta) + o(\|\boldsymbol{\beta}\|), \tag{152}$$

$$I^2_{N,r}/A^2_{N,r} \xrightarrow{\text{a.s.}} (\boldsymbol{\beta}'\mathbf{C}^{-1}\boldsymbol{\beta})\,\rho^2(\delta)I(f)/\nu(\delta) + o(\|\boldsymbol{\beta}\|). \tag{153}$$

Sub H_0 din (71), după teorema 8, $I^2_{N,r}/A^2_{N,r} \xrightarrow{R} U_1$ și deci se aplică (136).

Scanned with OKEN Scanner

Ținînd seama de teorema 8, (110), (120), (135) (pentru U_1) și (149) și (150), putem folosi rezultatele privind eficiența asimptotică relativă în sensul lui Bahadur (EARB).

EARB a testului de tip Kolmogorov-Smirnov relativă la testul terminal bazat pe $L_{N,r}$ este dată de

$$e_{K,T} = [\xi^*(\bar{F}^{-1}(\delta))]^2/\xi(\bar{F}^{-1}(\delta)) =$$

(154)

$$= \{\sup_{-\infty<x<\bar{F}^{-1}(\delta)} [\xi(x)]\}/\xi(\bar{F}^{-1}(\delta)) \geqslant 1$$

unde avem egalitate (printre alte cazuri) cînd $\xi(x)$ este nedescrescătoare în x.

EARB a testului de tip Cramer-von Mises în raport cu testul terminal bazat pe $L_{N,r}$ este

$$e_{M,T} = [\xi^0(\bar{F}^{-1}(\delta))/\xi(\bar{F}^{-1}(\delta))]\,(\pi^2/4) =$$

(155)

$$= \left[\left\{\int_{-\infty}^{\bar{F}^{-1}(\delta)} [\xi(y)/\xi(\bar{F}^{-1}(\delta))]\,\mathrm{d}\nu(\bar{F}(y))\right\} \Big/ \nu(\delta)\right][\pi^2/4)].$$

EARB a testului de tip Kolmogorov-Smirnov în raport cu testul de tip Cramer-von Mises este dată de

$$e_{K,M} = (4/\pi^2))]\ \xi^*(\bar{F}^{-1}(\delta))]^2/\xi^0(\bar{F}^{-1}(\delta))) =$$

(156)

$$= (4/\pi^2)\left\{\sup_{-\infty<x<\bar{F}^{-1}(\delta)} \xi(x)/[\{\nu(\delta)\}^{-1}\int_{-\infty}^{\bar{F}^{-1}(\delta)} \xi(y)\,\mathrm{d}\nu(\bar{F}(y))\right]\Big\}.$$

Obișnuit, o margine inferioară pentru (156) este $4/\pi^2 = 0{,}4053$.

Ne ocupăm de alternativele locale pentru care au loc (151) — (153). În acest caz EARB limită se reduce la

(157) $$e^*_{M,T} = (\pi^2/4)\left(\int_0^\delta \rho^2(u)\,\mathrm{d}\nu(u)\right)\Big/[\nu(\delta)\rho^2(\delta)],$$

(158) $$e^*_{K,M} = (4/\pi^2)\,(\sup_{0<t<\delta}\rho^2(t)\nu(\delta))\Big/\left(\int_0^\delta \rho^2(u)\,\mathrm{d}\nu(u)\right)$$

(159) $$e^*_{K,T} = [\sup_{0<t<\delta}\rho^2(t)]/\rho^2(\delta) \geqslant 1.$$

Chatterjee, Sen (1973) au studiat optimalitatea lui $\varphi(u, f)$ definit de (XXXVI.119). Din (XXXVI.119), (124) și (125) urmează că pentru $\varphi \equiv \varphi(u, f)$ (pînă la o constantă scalar)

$$(160) \qquad \nu(t) = \rho(t) = \left(\int_0^1 [\varphi_t(u, f)]^2 \, du \Big/ \int_0^1 [\varphi(u, f)]^2 \, du, \qquad 0 \leqslant t \leqslant 1.$$

astfel că din (157)—(160) avem

$$e^*_{M,T} = \pi^2/12 = 0{,}8225$$

$$e^*_{K,M} = 12/\pi^2 = 1{,}2159$$

$$e^*_{K,T} = 1$$

Observație. În acest caz sîntem înclinați să folosim teste de tip Kolmogorov-Smirnov. Totuși, aspectul este diferit cînd $\varphi \neq \varphi(u, f)$. De exemplu, presupunem că se folosește scorul exponențial

$$\varphi(u) = -1 - \log(1 - u), \qquad 0 < u < 1,$$

în timp ce repartiția este logistică. În acest caz $e^*_{K,M}$ se reduce la $15/2\pi^2 = 0{,}7599$ astfel că testele de tip Cramer-von Mises sînt mai bune decît cele de tip Kolmogorov-Smirnov.

„GH. ASACHI"
IAȘI

Scanned with OKEN Scanner

BIBLIOGRAFIE

1. ADICHIE, J.N., *Asymptotic efficiency of a class of nonparametric tests for regression parameters.* Ann. Math. Statist., *38*, 884–893 (1967 a).
2. ADICHIE, J.N., *Estimates of regression parameters based on rank test.* Ann. Math. Statist., *38*, 894–904 (1967 b).
3. ADICHIE, J.N., *Rank score comparison of several regression parameters.* Ann. Statist., *2*, 396–402 (1974).
4. ADICHIE, J.N., *Rank test of subhypotheses in the general linear regression.* Ann. Statist., *6*, 1012–1027 (1978).
5. AÎVAZIAN, S., *Étude statistique des dépendances.* Edit. Mir, Moscou, 1970.
6. ALBERT, A., *Fixed size confidence ellipsoids for linear regression parameters.* Ann. Math. Statist., *37*, 1602–1630 (1966).
7. ALLEN, D.M., *Mean square error of prediction as a criterion for selecting variables.* Technometrics, *13*, 469–475 (1971).
8. ALLEN, D.M., *The relationship between variable selection and data augmentation and a method for prediction.* Technometrics, *16*, 125–127 (1974).
9. ANDERSON, R.L., *Distribution of the serial correlation coefficient.* Ann. Math. Statist., *13*, 1–13 (1942).
10. ANDERSON, T.W., *An introduction to multivariate statistical analysis.* Wiley, New York, 1958.
11. ANDREWS, D.F., *A robust method for multiple linear regression.* Technometrics, *16*, 523–531 (1974).
12. ANDREWS, D.F., BICKEL, P.J., HAMPEL, F.R., HUBER, P.J., ROGERS, W.H., TUKEY, J.W., *Robust Estimates of Location : Survey and Advances.* Princeton University Press, Princeton, 1972.
13. ANSCOMBE, F. J., TUKEY, J.W., *The examination and analysis of residuals.* Technometries, *5*, 141–160 (1963).
14. BAKER, G.A., *The significance on the product-moment coefficient of correlation with special reference to the character of the marginal distributions.* J.A.S.A. *25*, 387–396 (1930).
15. BALAKRISHNAN. A.V., *An operator theoretic formulation of a class of control problems and a steepest descent method of solution.* Journal of Control, *1*, 109–127 (1963).
16. BANERJEE, D.P., *On the moments of the multiple correlation coefficient in samples form normal populations.* Journal of the Indian Society of Agricultural Statistics, *4*, 88–90 (1952).
17. BARON, T., *Metode statistice pentru analiza și controlul calității producției.* Edit. didactică și pedagogică, București, 1979.
18. BARTLETT, M.S., *Futher aspects of the theory of multiple regression.* Proc. Camb. Phil. Soc., *34*, 33–40 (1938).
19. BARTLETT, M.S., *The large sample theory of sequential tests.* Proc. Camb. Phil. Soc., *42*, 239–244 (1946).
20. BARTLETT, M.S., *Multivariate analysis.* J.R.S.S. Suppl. *9*, 176–190 (1947).
21. BELL, C.B., *Mutual information and maximal correlation as measures of dependence.* Ann. Math. Statist., *33*, 587–595 (1962).
22. BERNSTEIN, S., *Fondements géométriques de la théorie des corrélations.* Metron., **7** (2)3, 1928.
23. BICKEL, P.J., *On some robust estimates of location.* Ann. Math. Statist., *36*, 847–858 (1965).
24. BICKEL, P.J., *On some analogues to linear combinations of order statistics in the linear model.* Ann. Statist., *1*, 597–616 (1973).
25. BICKEL, P.J., *One-step Huber estimates in the linear model* J.A.S.A., *70*, 428–433 (1975).
26. BICKEL, P.J., *Another look at robustness : A reviews and some new developments (reply to discussant).* Scand. J. Statist., *3*, 167 (1976).
27. BIRNBAUM, A., *Sequential test for variance ratio and components of variance.* Ann. Math. Statist., *29*, 504–514 (1958).
28. BIRNBAUM, Z.W., *Effect of linear truncation on a multinormal population.* Ann. Math. Statist., *21*, 272–279 (1950).

Scanned with OKEN Scanner

130. GREVILLE, T.N.E., *Some applications of the pseudoinverse of a matrix.* SIAM Review, *2*, 15—22 (1960).
131. GROSS, ALAN, M., *Confidence interval robustness with longtailed symmetric distribution* J.A.S.A., *71*, 409—416 (1976).
132. GUIAŞU, S., *Aplicaţii ale teoriei informaţiei. Sisteme dinamice. Sisteme cibernetice.* Edit. Academiei R.S.R., Bucureşti, 1968.
133. GUIAŞU, S., *Mathematical structure of finite random cybernetic systems.* Springer Verlag, Wien, 1971.
134. GUIAŞU, S., *Information theory with new applications.* Mc Graw-Hill Book Co., New York—Düsseldorf, 1977.
135. GUIAŞU, S., MALIŢA, M., *Triade*, Edit. Ştiinţifică, Bucureşti, 1973.
136. GUILFORD, J.P., LYONS, T.C., *On determining the reliability and significance of a tetrachoric coefficient of correlation.* Psychometrika, *7*, 243, 1942.
137. GURLAND, J., *A relatively simple form of the multiple correlation coefficient.* J.R.S.S., **B**, *30*, 276—283 (1968).
138. GURLAND, J., MELTON, R.C., *Further consideration of the distribution of the multiple correlation coefficient.* J.R.S.S., *32* (1970).
139. HÁJEK, J., *Some extensions of the Wald-Wolfowitz-Noether theorem.* Ann. Math. Statist., *32*, 506—523 (1961).
140. HÁJEK, J., *Asymptotically most powerful rank order tests.* Ann. Math. Statist., *33*, 1127—1147 (1962).
141. HÁJEK, J., *Asymptotic normality of simple linear statistics under alternatives.* Ann. Math. Statist., *39*, 325—346 (1968).
142. HÁJEK, J., SIDAK, Z., *Theory of Rank Statistics.* Academic Press, New York, 1967.
143. HALDANE, J.B.S., *Moments on the distributions of powers and products of normal variates.* Biometrika, 32, 226—242 (1942).
144. HAMDAN, M.A., *The equivalence of tetrachoric and maximum likelihood estimates of* ρ *in* 2×2 *tables.* Biometrika, *57*, 212—215 (1970).
145. HAMILTON, M., *Nomogram for the tetrachoric correlation coefficient.* Psychometrika, *13*, 259 (1948).
146. HAMPEL, F.R., *A general qualitative definition of robustness.* Ann. Math. Statist., *42*, 1887—1896 (1971).
147. HAMPEL, FRANK, R., *The influence curve and its role in robust estimation.* J.A.S.A., *69*, 383—393 (1974).
148. HANNAN, E.J., *The asymptotic power of tests based upon multiple correlation.* J.R.S.S., **B**, *18*, 227—233 (1956).
149. HARLEY, BETTY, I., *A note on the probability integral of the correlation coefficient.* Biometrika, *41*, 278—280 (1954).
150. HARLEY, BETTY, I., *Some properties of an angular transformation for the correlation coefficient.* Biometrika, *43*, 219—224 (1956).
151. HARLEY, BETTY, I., *Relation between the distribution of noncentral t and of a transformed correlation coefficient.* Biometrika, *44*, 219—224 (1957).
152. HAYES, S.P.JR., *Tables of the standard error of the tetrachoric correlation coefficient.* Psychometrika, *8*, 193—203 (1943).
153. HAYES, S.P.JR., *Diagrams for computing tetrachoric correlation coefficients from percentage differences.* Psychometrika, *11*, 163 (1946).
154. HEMMERLE, W.J., *An explicit solution for generalized ridge regression.* Technometrics, *17*, 309—313 (1975).
155. HETTMANSPERGER, T.P., MCKEAN, J.W., *Computational problems involved in analysis of linear models based on ranks.* Proc. Statist. Comp. Sect. 1976. Amer. Statist. Assoc. Meetings. Washington, D.C. American Statistical Association.
156. HETTMANSPERGER, T.P. MEKEAN, J.W., *A robust alternative based on ranks to least squared in analyzing linear models.* Technometrics, *19*, 275—284 (1977).
157. HETTMANSPERGER, T.P., MCKEAN, J.W., *Statistical inferance based on ranks.* Psychometrika, *43*, 69—79 (1978).
158. HILGERS, J.W., *On the equivalence of regularization and certain reproducing Hilbert space approaches for solving first kind problems.* SIAM J. Num. Anal., *13*, 172—184 (1976).
159. HILL, R.W., HOLLAND, P.W., *Two robust alternatives to least-squares regression.* J.A.S.A., *72*, 828—833 (1977).
160. HIRSCHFELD, H.O., (HARTLEY, H.O.) *The distribution of the ratio of covariance estimates in samples drawn from normal bivariate populations.* Biometrika, *29*, 65—79 (1937).
161. HOADLEY, B., A *Bayesian look at inverse linear regression.* J.A.S.A., *65*, 356—369 (1970).
162. HOCKING, R.R., LESLIE, R.N., *Selection of the best subset in regression analysis.* Technometrics, *9*, 531—540 (1967).
163. HODGES, J.L., LEHMANN, E.L., *Estimation for location based on rank tests.* Ann. Math. Statist., *34*, 598—611 (1963).

Scanned with OKEN Scanner

164. Hodgson, *On the sampling distribution of the multiple correlation coefficient.* Prezentată la adunarea I.M.S. din Washington, D.C., Decembrie 1967.
165. Hoeffding, W., *Optimum nonparametric tests.* Proc. Second Berkeley Symp. Mat. Statist. Prob., *1*, 82—92, 1951.
166. Hoerl, A.E., *Application of ridge analysis to regression problems.* Chemical Engineering Progress, *58*, 54—59 (1962).
167. Hoerl, A.E., *Ridge analysis.* Chemical Engineering Progress Symposium Series, *60*, 67—77 (1964).
168. Hoerl, A.E., Kennard, R.W., *Ridge regression: Biased estimation for nonorthogonal.* Technometrics, *12*, 55—69 (1970 a).
169. Hoerl, A.E., Kennard, R.W., *Ridge regression: Applications to nonorthogonal problems.* Technometrics, *12*, 69—83 (1970 b).
170. Hoerl, A.E., Kennard, R.W. *Ridge regression: iterative estimation of the biasing parameter.* Comm. Statist., *A5*, 77—88 (1976).
171. Hoerl, A.E., Kennard, R.W., Baldwin, K.F. *Ridge regression: some simulations.* Comm. Statist., *4*, 105—123 (1975).
172. Hogg, Robert, V., *Adaptive robust procedures: A partial review and some suggestions for future applications and theory.* J.A.S.A., *69*, 909—927 (1974).
173. Hoog, Robert, V., *Statistical robustness: one view of its use in applications today.* Amer. Statist., *33*, 108—115 (1979).
174. Holland, P.W., Welsch, R.E., *Robust regression using iteratively reweighted least-squares.* Comm. Statist., *A*, *6*, 813—827 (1977).
175. Hollander, M., *A distribution free test for parallelism.* J.A.S.A., *65*, 387—394 (1970).
176. Honsenholder, A., *The theory of matrices in numerical analysis.* Blaisdell, New York, 1964.
177. Hooper, J.W., *Simultaneous equations and canonical correlation theory.* Econometrica, *27*, 245—256 (1959).
178. Hotelling, H., *New light on the correlation coefficient and its transforms.* J.R.S.S., **B**, *15*, 193—224 (Discuție 225—232) (1953).
179. Hsu, P.L., *On the limiting distribution of the canonical correlations.* Biometrika, *32*, 38—45 (1941).
180. Hsu, P.L., *On the asymptotic distributions of certain statistics used in testing the independence between successive observations from a normal population.* Ann. Math. Statist., *17*, 350—354 (1946).
181. Huber, P.J., *Robust estimation of a location parameter.* Ann. Math. Statist., *35*, 73—101 (1964).
182. Huber, P.J., *Robust confidence limits.* Z. Wahrscheinlichkeitstheorie und Verw. Gebeite, *10*, 269—278 (1968).
183. Huber, P.J., *Studentizing robust estimates:* In *Nonparametric techniques in statistical inference.* Ed. M.L. Puri. Cambridge University Press, London, 435—463, 1970.
184. Huber, P.J., *Robust statistics. A review.* Ann. Math. Statist., *43*, 1041—1067 (1972).
185. Huber, P.J. *Robust regression: Asymptotics, conjectures and Monte Carlo.* Ann. Statist., 799—822 (1973).
186. Iliescu, D.V., Vodă, V. Gh., *O metodă de reducere a numărului mediu de probe în controlul de recepție a produselor.* Revista de statistică, anul XXI, *12*, 80—87 (1972).
187. Ionescu, H., *Statistică matematică.* Edit. didactică și pedagogică, București, 1962.
188. Iosifescu, M., Mihoc, Gh., Theodorescu, R., *Teoria probabilităților și statistică matematică.* Edit. tehnică, București, 1966.
189. Isserlis, L., *On certain probable errors and correlation coefficients of multiple frequency distributions with skew regression.* Biometrika, *11*, 185—190 (1915).
190. Ito, K., *On the effect of teteroscedasticity and nonnormality upon some multivariate test procedures.* In *Multivariate Analysis*, 2. Ed. P.R. Krishnaiah. 87—120. Academic Press, New York, 1969.
191. Jaeckel, L.A., *Robust estimates of location: Symmetry and asymmetry contamination.* Ann. Math. Statist., *42*, 1020—1034 (1971).
192. Jaeckel, L.A., *Estimating regression coefficients by minimizing the dispersion of the residuals.* Ann. Math. Statist., *43*, 1449—1458 (1972).
193. Jaiswal, M.C., Khatri, C.G., *Estimation of parameters for the selected samples from bivariate normal populations.* Metron., *26*, 1—8 (1967).
194. James, A.T, *Distributions of matrix variates and latent roots derived from normal.* Ann. Math. Statist., *35*, 475—501 (1964).
195. James, W., Stein, C.M., *Estimation with quadratic loss.* Proc. 4th. Berkeley Simp., *1*, 361—379 (1961).
196. Jenkins, G.M., *An angular transformation for the serial correlation coefficient.* Biometrika, *41*, 261—265 (1954).

Scanned with OKEN Scanner

197. JENKINS, G.M., *Tests of hypotheses in the linear autoregressive model* I. Biometrika, *41*, 405–419 (1954).
198. JENKINS, G.M., *Tests of hypotheses in the linear autoregressive model*. II. Biometrika, *43*, 186–199 (1956).
199. JENKINS, W.L., *An improved method for tetrachoric*. Psychometrika, *20*, 253 (1955).
200. JOHNSON, N.L., *Some notes on the application of sequential methods in the analysis of variance*. Ann. Math. Statist., *24*, 614–623 (1953).
201. JOHNSON, N.L., *Sequential procedures in certain component of variance problems*. Ann. Math. Statist. 25, 357–366 (1954).
202. JOHNSON, N.L., KOTZ, S., *Distributions in Statistics : Continuous Univariate Distributions* 2. Houghton Mifflin, Boston ; 1970.
203. JOHNSON, N.L., KOTZ, S., *Distributions in Statistics : Continuous Multivariate Distributions*. Wiley, New York, 1972.
204. JOHNSON, R.A., WEHRLY, T., *Measures and models for angular correlation and angular-linear correlation* J.R.S., B, *39*, 222–229 (1977).
205. JOHNSON, R.A., WEHRLY, T., *Some angular-linear distributions and related regression models* J.A.S.A., *73*, 602–606 (1978).
206. JUPP, P.E., MARDIA, K.V., *A general correlation coefficient for directional regression problems*. Biometrika, *67*, 163–173 (1980).
207. JUREČKOVÁ, J., *Asymptotic linearity of a rank statistic*. Ann. Math. Statist., *40*, 1889–1900 (1969).
208. JUREČKOVÁ, J., *Nonparametric estimate of regression coefficients* Ann. Math. Statist., *42*, 1328–1338 (1971 a).
209. JUREČKOVÁ, J., *Asymptotic independence of rank test statistic for testing symmetry on regression*. Sankhya, A, *33*, 1–18 (1971 b).
210. JUREČKOVÁ, J., *Asymptotic relations of M-estimates and R-estimates in linear regression model*. Ann. Statist., *5*, 464–472 (1977).
211. KAMAT, A.R., *Incomplete and absolute moments of the multivariate normal distribution with some applications*. Biometrika, *40*, 20–34 (1953).
212. KAMAT, A.R., *Hypergeometric expansions for incomplete moments of the bivariate normal distribution*. Sankhya, *20*, 317–320 (1958).
213. KAMAT, A.R., *Incomplete moments of the trivariate normal distribution*. Sankhyā, *20*, 321–322 (1958).
214. KAMAT, A.R., SATHE, Y.S., *Asymptotic power of certain test criteria (based on first and second differences) for serial correlation between succesive observations*. Ann. *Math. Statist.*, *33*, 186–200 (1962).
215. KARIYA, T., EATON, M.L., *Robust tests for spherical symmetry*. Ann. Statist., *5*, 206–215 (1977).
216. KELKER, D., *Distribution theory of spherical distributions and a location-scale parameter generalization*. Sankhyā, A, *32*, 419–430 (1970).
217. KEMPTHORNE, O., *The Design and Analysis of Experiments*. Wiley, New York, 1952.
218. KENDALL, M.G., *Proof of relations connected with the tetrachoric series and its generalization*. Biometrika, *32*, 196–198 (1941).
219. KENDALL, M.G., *Rank correlation methods*. London, 1948.
220. KENDALL, M.G., *Rank and product-moment correlation*. Biometrika, *36*, 177–193 (1949).
221. KENDALL, M.G., *Note on bias in the estimation of autocorrelation*. Biometrika, *41*, 403–404 (1954).
222. KENDALL, M.G., *The moments of the Leipnik distribution*. Biometrika, *44*, 270–272 (1957).
223. KENDALL, M.G., *A course in multivariate analysis*. Hafner Publishing Co., New York, 1957.
224. KENDALL, M.G., *The evergreen correlation coefficient* [p. 274–277] in *Essays in Honor of Harold Hotelling* (Ed. I. Olkin), Stanford University Press 1960.
225. KENDARLL, M.G., STUART. A., *The Advanced Theory of Statistics*, vol. I. Distribution Theory. Griffin, London, 1958.
226. KENDALL, M.G., STUART, A., *The Advanced Theory of Statistics*. vol. II. *Inference and Relationship*. Griffin, London, 1961.
227. KHATRI, C.G., *A note on a large sample distribution of a transformed multiple correlation coefficient*. Ann. Inst. Statist., Math., Tokyo, *18*, 375–380 (1966).
228. KHATRI, C.G., *Some results for the singular normal multivariate regression models*. Sankhyā, A, *30*, 267–280 (1968).
229. KHATRI, C.G., MARDIA, K.V., *The von Mises-Fisher matrix distribution in orientation statistics*. J.R.S.S., B, *39*, 95–106 (1977).
230. KONISHI, SADANORI, *An approximation to the distribution of the sample correlation coefficient*. Biometrika, *65*, 654–658 (1978).
231. KOOPMANS, T.C., *Serial correlation and quadriatic forms in normal variables*. Ann. Math. Statist., *13*, 14–33, (1942).

Scanned with OKEN Scanner

232. KOUL, H.L., *Asymptotic behavior of Wilcoxon type confidence regions in multiple linear regression*. Ann. Math. Statist., *40*, 1950–1979 (1969).
233. KOUL, H.L., *A class of ADF tests for subhypothesis in the multiple linear regression*. Ann. Math. Statist. *41*, 1273–1281 (1970).
234. KOUL, H.L., *Asymptotic behavior of a class of confidence regions based on ranks in regression*. Ann. Math. Statist., *42*, 466–476 (1971).
235. KRAEMER, H.C., *Improved approximation to the non-null distribution of the correlation coefficient*. J.A.S.A., *68*, 1004–1008 (1973).
236. KRAEMER, HELEN, CH., *The non-null distribution of the Spearman rank correlation coefficient*. J.A.S.A., *69*, 114–117 (1974).
237. KRAFT, C.H., VAN EEDEN, C., *Linearised rank estimates and signed-rank estimates for the general linear hypothesis*. Ann. Math. Statist., *43*, 42–57 (1972).
238. KRAMER, K.H., *Tables for constructing confidence limits on the multiple correlation coefficient*. J.A.S.A., *58*, 1082–1085 (1963).
239. KRUTCHKOFF, R.G., *Clasical and inverse regression methods of calibration*. Technometrics, *9*, 425–439 (1967).
240. KULLBACH, S., *A note on the multiple correlation coefficient*. Metron, *12*, 67–72 (1936).
241. LACHENBRUCH, P.A., SNEERINGER, C., REVO, L.T., *Robustness of the linear and quadratic discriminant function to certain types of non-normality*. Comm. Statist., *1*, 39–56 (1973).
242. LA MOTTE, L.R., HOCKING, R.R., *Computational efficiency in the selection regression variables*. Technometrics, *12*, 83–94 (1970).
243. LARSEN, W.A., MCCLEARY S.J., *The use of partial residual plots in regression anaylsis* Technometrics, *14*, 781–790. (1972).
244. LAWLESS, J.F., SINGHAL, K., *Efficient screening of non-normal regression models*. Biometrika, *34*, 318–327 (1978).
245. LAWLESS, J.F., WANG, P., *A simulation study of ridge and other regression estimates*. Comm. Statist., **A**, *5*, 307–324 (1976).
246. LAWLEY, D.N., *Tests of significance in canonical analysis*. Biometrika, *46*, 59–66 (1969).
247. LEHMANN, E.L., *Testing Statistical Hypothesis*. Wiley, New York, 1959.
248. LEHMANN, E.L., *Some concepts of dependence*. Ann. Math. Statist. *37*, 1137–1153 (1966).
249. LEIPNIK, R.B., *Distribution of the serial correlation coefficient in a circulary correlated universe*. Ann. Math. Statist., *18*, 80–87 (1947).
250. LEIPNIK, R.B., *Note on the characteristic function of a serial-correlation distribution*. Biometrika, *45*, 559–562 (1958).
251. LIPOW, M., EIDEMILLER, R., L. *Application of the bivariate normal distribution to a stress vs strength problem in reliability analysis*. Technometrics, *6*, 325–328 (1964).
252. LOYNES, R.M., *On idempotent matrices*. Ann. Math. Statist., *37*, 295–296 (1966).
253. LOYNES, R.M., *On Cox and Snell's general definition of residuals*. J.R.S.S., **B**, *31*, 103–106 (1969).
254. MACKENZIE, J.K., *The estimation of an orientation relationship*. Acta Cryst., *10*, 61–62 (1957).
255. MADOW, W.G., *Note on the distribution of the serial correlation coefficient*. Ann. Math. Statist., *16*, 308–310 (1945).
256. MAHALANOBIS, P.C., BOSE, R.C., ROY, S.N., *Normalization of statistical variates and the use of rectangular coordinates in the theory of sampling distributions*. Sankhyā, *3*, 1–40 (1937).
257. MAJUMDAR, H., SEN, P.K., *Nonparametric tests for multiple regression under progressive censoring*. J, Multivar. Anal., *8*, 73–95 (1978).
258. MARDIA, K.V., *Measures of multivariate skewness and kurtosis with applications*. Biometrika, *57*, 519–530 (1970).
259. MARDIA, K.V., *Statistical of Directional Data*. Academic Press, London, 1972.
260. MARDIA, K.V., *Statistics of directional data* (with discussion) J.R.S.S., **B**, *37*, 349–393 (1975).
261. MARDIA, K.V., *Linear-circular correlation coefficients and rhythmometry*. Biometrika, *63*, 403–405 (1976).
262. MARDIA, K.V., PURI, M.L., *A spherical correlation coefficient robust agains scale*. Biometrika, *65*, 391–395 (1978).
263. MARDIA, K.V., SPURR, B.D., *Multisample tests for multimodal and axial circular populations*. J.R.S.S., **B**, *35*, 422–436 (1973).
264. MARDIA, K.V., SUTTON, T.W., *A model for cylindrical variables with applications*. J.R.S.S. **B**, *40*, 229–233 (1978).
265. MARINESCU, I., MOIGEANU, C., NICULESCU, R., RANCU, N., URSEANU, V., *Elemente de statistică matematică și aplicațiile ei*. Edit. științifică, București, 1966.
266. MARONNA, R.A., *Robust M-estimators of location and scatter*. Ann. Statist., *4*, 51–67 (1976).

Scanned with OKEN Scanner

267. MARQUARDT, D.W., *An algorithm for least-squares estimation of nonlinear parameters.* J. Soc. Indust. Appl. Math., *11*, 431—441 (1963).
268. MARQUARDT, D.W., *Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation.* Technometrics, *12*, 591—612 (1970).
269. MARRIOTT, F.H.C., POPE, J.A., *Bias in the estimation of autocorrelation.* Biometrika, *41*, 390—402 (1954).
270. MCDONALD, G., GALARNEAU, D., *A Monte Carlo evaluation of some ridge type estimators.* J.A.S.A., *70*, 407—416 (1975).
271. MCKEAN, J.W., HETTMANSPERGER, T.P. *Tests of hypothesis based on ranks in the general linear model.* Comm. Statist., A, 5, 693—709 (1976).
272. MCKEAN, J.W., HETTMANSPERGER, T.P., *A robust analysis of the general linear model based on one step R-estimates.* Biometrika, *65*, 571—579 (1978).
273. MEHTA, J.S., GURLAND, J., *Some properties and an application of a statistic arising in testing correlation.* Ann. Math. Statist., *40*, 1736—1745 (1969).
274. MIHĂILĂ, N., *Statistică matematică*, vol. I, II. Litografia A.S.E., București, 1972.
275. MIHĂILĂ, N., POPESCU, O., *Matematici speciale aplicate în economie.* Edit. didactică și pedagogică, București, 1978.
276. MIHOC, GH., *Cîteva considerații asupra energiei și corelației informaționale.* În Studii de statistică informațională. Direcția de statistică, 1978.
277. MIHOC, GH., CIUCU, G., CRĂIU, V., *Teoria probabilităților și statistică matematică.* Edit. didactică și pedagogică, București, 1970.
278. MIHOC, GH., FIRESCU, D., *Statistică matematică.* Edit. didactică și pedagogică, București, 1966.
279. MIHOC, GH., URSEANU, V., *Matematici aplicate în statistică.* Edit. Academiei, 1962.
280. MIHOC, GH., URSEANU V., *Sondaje și estimații statistice.* Teorie și aplicații. Edit. tehnică, București, 1977.
281. MORAN, P.A.P., *Rank correlation and product-moment correlation.* Biometrika, *35*, 203—206. (1948).
282. MORAN, P.A.P., *The distribution of the multiple correlation coefficient.* Proc. Camb. Phil., Soc., *46*, 521—522 (1950).
283. MUIRHEAD, R.J., *Latent roots and matrix variates : A review of some asymptotic results.* Ann. Statist., *6*, 5—33 (1978).
284. MUIRHEAD, R.J., WATERNAUX; C.M., *Asymptotic distribution in canonical correlation analysis and other multivariate procedures for nonnormal populations.* Biometrika, *67*, 31—43 (1980).
285. MURTHY, V.K., *On the distribution of the sum of circular serial correlation coefficients and the effect on non-normality on its distribution (abstract).* Ann. Math. Statist., *31*, 239—240 (1960).
286. NABEYA, S., *Note on the moments of the transformed correlation.* Ann. Inst. Statist. Math.. Tokyo, *3*, 1 (1951).
287. NABEYA, S., *Absolute moments in 2-dimensional normal distribution.* Ann. Inst. Statist. Math., Tokyo, *3*, 2—6 (1951).
288. NABEYA, S., *Absolute moments in 3-dimensional normal distribution.* Ann. Inst. Statist. Math., Tokyo, *4*, 15—20 (1952).
289. NADARAYA, E.A., *On nonparametric estimate of density functions and regression curves.* Theory Prob. Appl., *10*, 186—190 (1965).
290. NATH, G.B., *Estimation in truncated bivariate normal distributions.* Applied Statistics, *20*, 313—319 (1971).
291. NETER, J., WASSERMAN, W., *Applied Linear Statistical Models.* Homewood. Richard D. Irwin. Inc.
292. NEUMANN, J. VON, *Distribution of the ratio of the square successive difference to the variance.* Ann. Math. Statist., *12*, 367—395 (1941).
293. NICHOLSON, C., *The probability integral for two variables.* Biometrika, *33*, 59—72 (1943).
294. OBENCHAIN, R.L., *Ridge analysis following a preliminary test of the shrunken hypothesis.* Technometrics, *17*, 431—441 (1975).
295. OBENCHAIN, R.L., *Good and optimal ridge estimators.* Ann. Statist., *6*, 1111—1122 (1978).
296. OBREJA, GH., STOICA, D., VODĂ, V. GH., *Aplicații ale teoriei regresiei la determinarea creșterii în timp a dimensiunilor rulmenților.* Revista de statistică, anul XXI, nr. 1, 27—38 (1972).
297. OLKIN, I., TATE, R.F., *Multivariate correlation models with mixed discrete and continuous variables.* Ann. Math. Statist., *32*, 448—465 (1961).
298. ONICESCU, O., *Energia informațională*, St. cerc. math., *18*, 1419—1420 (1966).
299. ONICESCU, O., *Corelație informațională.* Rev. de statistici, *6*, (1972).
300. ONICESCU, O., MIHOC, GH., *Lecții de statistică matematică.* Edit. tehnică, București, 1958.

Scanned with OKEN Scanner

395. SHEPPARD, W.F., *On the calculation of the double integral expressing normal correlation.* Trans. Cambridge Phil. Soc., *19*, 23—66 (1900).
396. SIEVERS, G.L., *Weighted rank statistics for simple linear regression.* J.A.S.A., 73, 628—631 (1978).
397. SIEVERS, G.L., *A weighted dispersion function for estimation in linear models.* Technical reports No. 62. Dep. of. Math. Western Michigan University 1979.
398. SLATER, L.J., *Confluent Hypergeometric Functions.* Cambridge Univ., Press, 1960.
399. SOMERVILLE, P.N., *Some problems of optimum sampling.* Biometrika, *41*, 420—429 (1954)
400. SOPER, H.E., *On the probable error of the correlation coefficient to a second approximation.* Biometrika, *9*, 91—115 (1913).
401. SOPER, H.E., *On the probable error for the biserial expression for the correlation coeffficient.* Biometrika, *10*, 384—390. (1915).
402. SOPER, H.E., *The general sampling distribution of the multiple correlation coefficient.* J.R.S.S., **A**, *92*, 445—447 (1929).
403. SOPER, H.E., YOUNG, A.W., CAVE, B.M., LEE, ALICE., PEARSON, K., *On the distribution of the correlation coefficient in small samples.* Apendice II la lucrările lui „Student" și R.A. Fisher. Biometrika, *11*, 328—413 1915—1917.
404. SPEARMEN, C., *The proof and measurement of association between two things.* Amer. J. Psychol. 15, 88 (1904).
405. SRIVASTAVA, M.S., *On fixed-width confidence bounds for regression parameters and mean vector.* J.R.S.S., **B**. *29*, 132—141 (1967).
406. SRIVASTAVA, M.S., *On fixed-width confidence bounds for regression parameters.* Ann. Math. Statist., *42*, 1403—1411 (1971).
407. STAMMBERGER, A., *Ein Nomogramm zur Beurteilung von Korrelationskoeffizienten.* Biometrische Zeitschrift., *10*, 80—83 (1968).
408. STECK, G.P., *A table for computing trivariate normal probabilities.* Ann. Math. Statist., *29*, 780—800 (1958).
409. STEIN, C.M., *Multiple regression.* Cap. 37 din *Essays in Honor of Harold Hotelling.* Stanford University Press, 1960.
410. STEIN, C.M., *Confidence sets for the mean of a multivariate normal distribution.* J.R.S.S., **B**. *24*, 265—296 (1962).
411. STEPHENS, M.A. *Vector correlation.* Biometrika, *66*, 41—48 (1979).
412. STOUT, W.F., *Some results on the complete and almost sure convergence of linear combinations of independent random variables and martingale differences.* Ann. Math. Statist., *39*, 1549—1562 (1968).
413. STOUT, W.F., *A martingale analogue of Kolmogorov's law of iterated logarithm.* Zeit. Wahr. Verw. Geb. *15*, 279—290 (1970).
414. STRASSEN, V., *Almost sure behaviour of sums of independent random variables and martingales.* Proc. Fifth. Berkeley Symp. Math. Statist. Prob. *2*, 315—343 (Ed. J. Neyman, L. LeCam) Univ. Calif. Press.
415. SUGIURA, N., *Derivatives of the characteristic root of a symmetric or a Hermitian matrix with two applications in multivariate analysis.* Comm. Statist., *1*, 393—417 (1973).
416. TALLIS, G.M., *Elliptical and radial truncation in normal populations.* Ann. Math. Statist., *34*, 940—944 (1963).
417. TATE, E.F., *The theory of correlation between two continuous variables when one variable is dichotomized.* Biometrika, *42*, 205—216 (1955).
418. TEICHROEW, D., *Probabilities associated with order statistics in sample from two normal populations with equal variances.* Chemical Corps Engineering Agency, Army Chemical Center Maryland (1955).
419. TERRY, M.E., *Some rank order tests which are most powerful against specific parametric alternatives.* Ann. Math. Statist., *23*, 346—366. (1952).
420. THEILER, G., TÖVISSI, L., *Entropia, eventual alte măsuri cu suport probabilistic în evaluarea unor informații economice.* Rev. de statistică, 10 (1976).
421. THEILER, G., TÖVISSI, L., *Entropia și alte măsuri de același gen în evaluarea informației economice.* Rev. de statistică 2 (1977).
422. THEOBALD, C.M., *Generalizations of mean squared error applied to ridge regression.* J.R.S.S., *B. 36*, 103—106 (1974).
423. THIEL, H., *A simplification of the BLUS procedure for analyzing regression disturbances.* J A.S.A., *63*, 242—251 (1968).

Scanned with OKEN Scanner